W4111
Introduction to Databases
Fall 2024

Computer Science Department

Columbia University

Welcome!

Eugene Wu

B.S.	U.C. Berkeley
Ph.D.	MIT
PostDoc	U.C. Berkeley
Professor	Columbia since Fall 2015

DBMSes, visualization, data analysis, cleaning, ML data systems.

www.eugenewu.net
ewu@cs.columbia.edu
421 Mudd

Office hours
Tues 10:15-11:15AM in person
or by appointment

Agenda

Overview

Course Overview

The Future of AI: How Artificial Intelligence Will Change the World

AI is constantly changing our world. Here are just a few ways AI will influence our lives.

Written by **Mike Thomas**

PEW RESEARCH CENTER | DECEMBER 10, 2018

Artificial Intelligence and the Future of Humans

Experts say the rise of artificial intelligence will make most people better off over the next decade, but many have concerns about how advances in AI will affect what it means to be human, to be productive and to exercise free will

BY JANNA ANDERSON AND LEE RAINIE

The Economist

Schools brief | Applications for AI

LLMs will transform medicine, media and more

But not without a helping (human) hand

EU Parliament
AI in the Digital Age (2021)

Notes that the world stands on the verge of the fourth industrial revolution; points out that in comparison with the three previous waves, initiated by the introduction of steam, electricity, and then computers, the fourth wave draws its energy from an abundance of data combined with powerful algorithms; stresses that today's digital revolution is shaped by its unprecedented scale, fast convergence, and the enormous impact of emerging technological breakthroughs on states, economies and societies;

Argues that artificial intelligence (AI) is the key emerging technology within the fourth industrial revolution; notes that AI is the control centre of the new data layer that surrounds us and which can be thought of as the fifth element after air, earth, water and fire; states that by 2030, AI is expected to contribute more than EUR 11 billion to the global economy, an amount that almost matches China's GDP in 2020;

Conventional View of AI/Data Science

Lone data scientist uses a static, clean table,
applies statistics or fits an ML model
to increase a well-defined score

See school, ML articles, Kaggle competitions, etc

Conventional View of AI/Data Science

Lone data scientist uses a static, clean table,
applies statistics or fits an ML model
to increase a well-defined score

See school, ML articles, Kaggle competitions, etc

Conventional View of AI/Data Science

Team

Dynamic Messy Non-tabular

Lone data scientist uses a static, clean table,

applies statistics or fits an ML model

Unavailable

to increase a well-defined score

unclear, ill-defined

Huge amount of "unseen labor" (data engineering)

in order to support real-world data science & ML

Ask an ML Ops Engineer

An on-call engineer's biggest nightmare

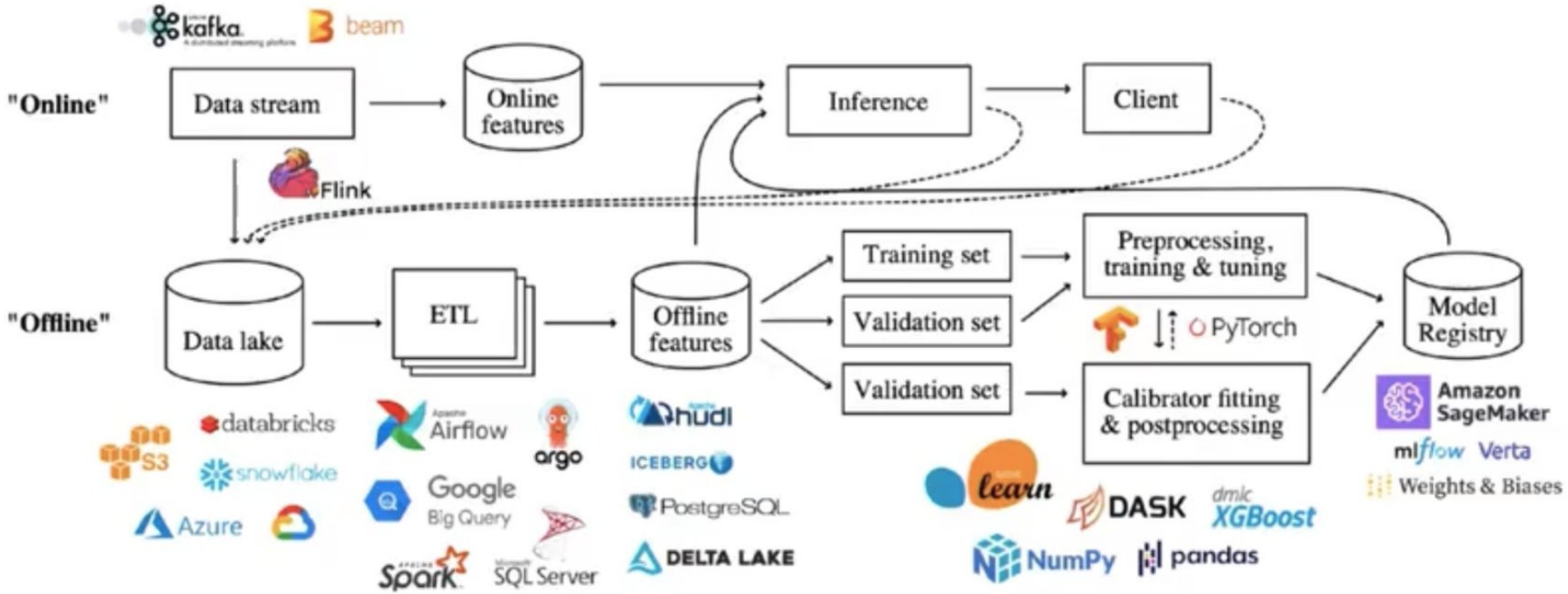
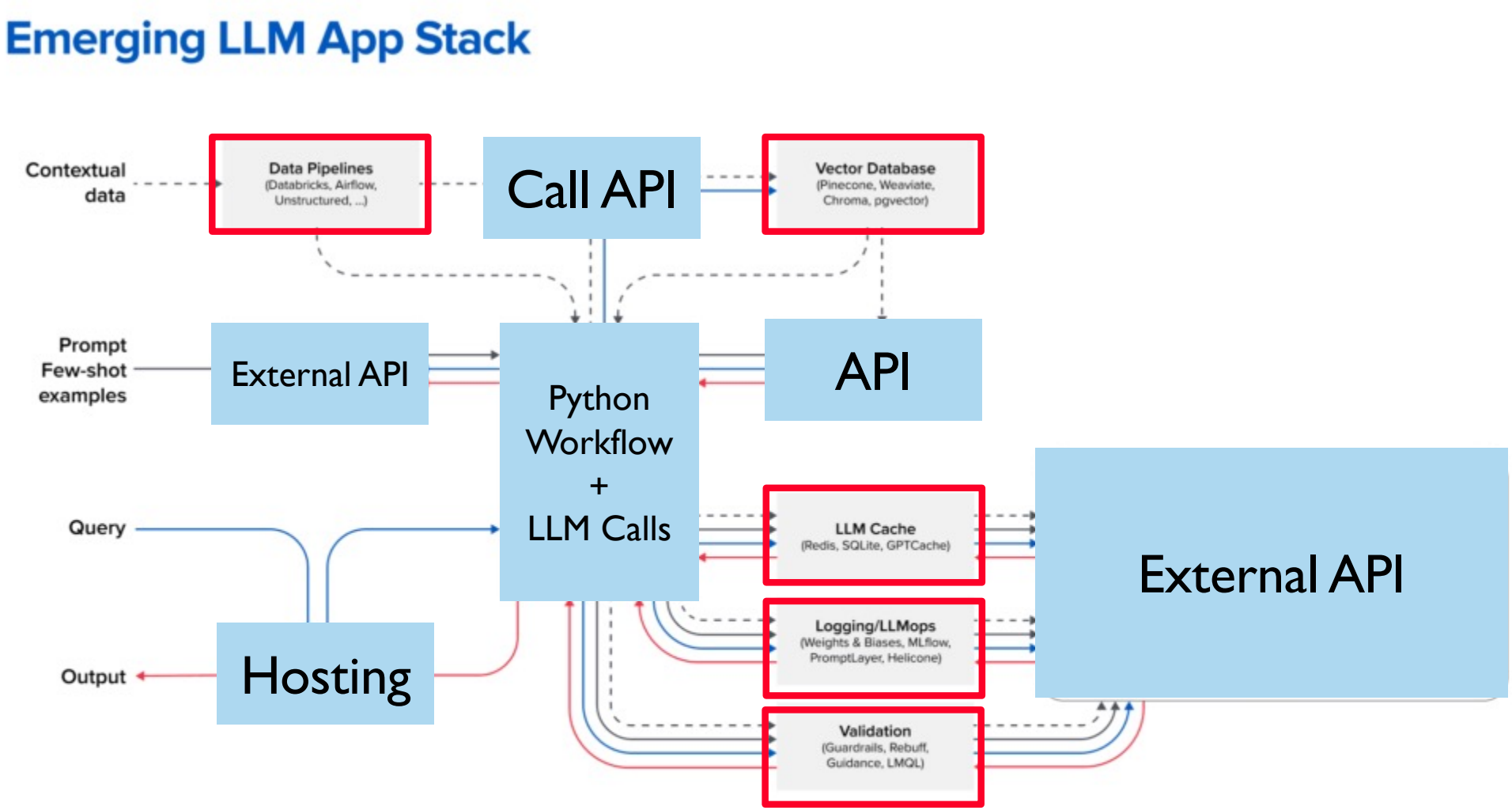


Figure 1: High-level architecture of a generic end-to-end machine learning pipeline. Logos represent a sample of tools used to construct components of the pipeline, illustrating heterogeneity in the tool stack. *Shankar et al. 2021*

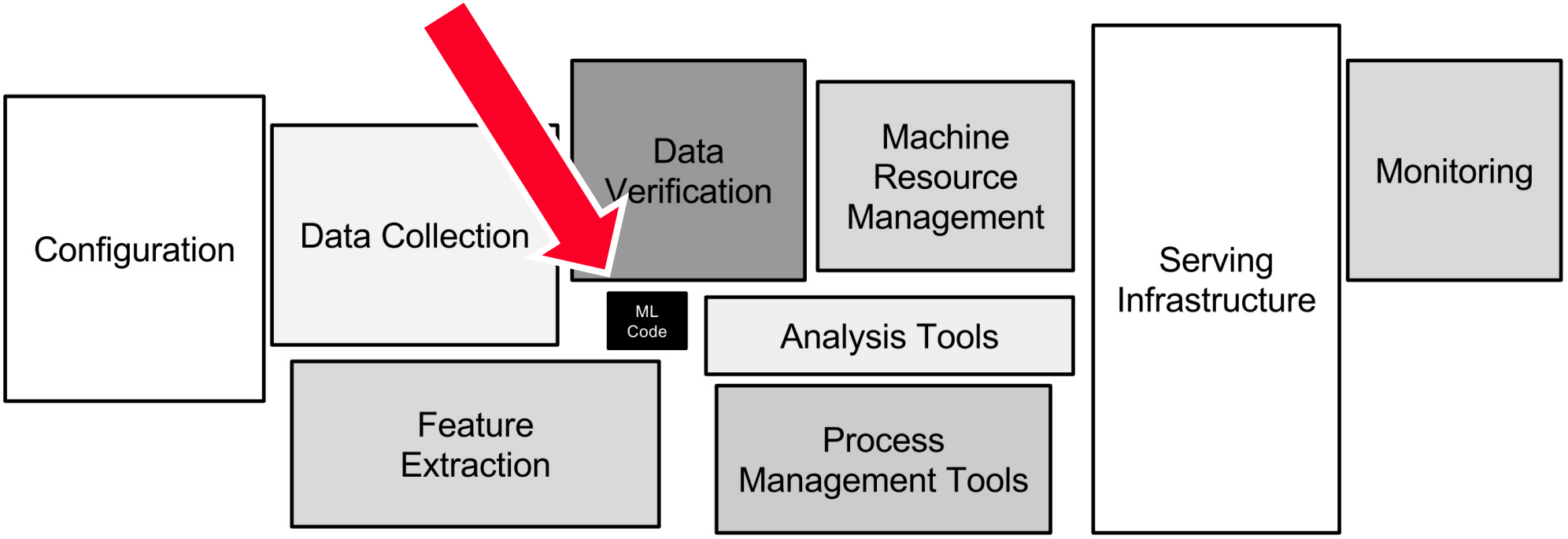
https://www.facebook.com/Engineering/videos/1578607659138164/

Or ask A16Z…


Emerging LLM App Stack
Contextual data
Data Pipelines
(Databricks, Airflow, Unstructured, ...)
Call API
Vector Database
(Pinecone, Weaviate, Chroma, pgvector)
Prompt
Few-shot examples
External API
Python Workflow + LLM Calls
API
Query
Output
Hosting
LLM Cache
(Redis, SQLite, GPTCache)
Logging/LLMops
(Weights & Biases, MLflow, PromptLayer, Helicone)
Validation
(Guardrails, Rebuff, Guidance, LMQL)
External API

Or ask Google…

Hidden Technical Debt in Machine Learning Systems



In Reality…

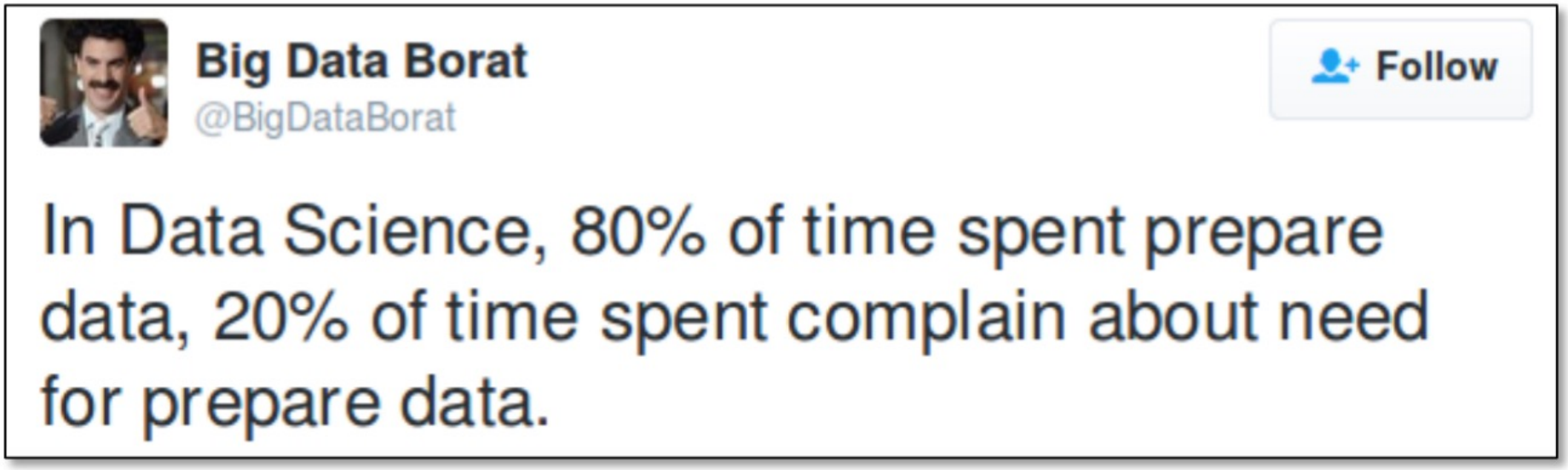
Data engineering dominates data science projects

Data engineer work >> data scientist work

Data engineering key to ML/AI/data science

Data management principles underly data engineering

Data Eng Dominates Data Science Projects



Most time spent on data engineering

Not viewed as sexy, but arguably most important

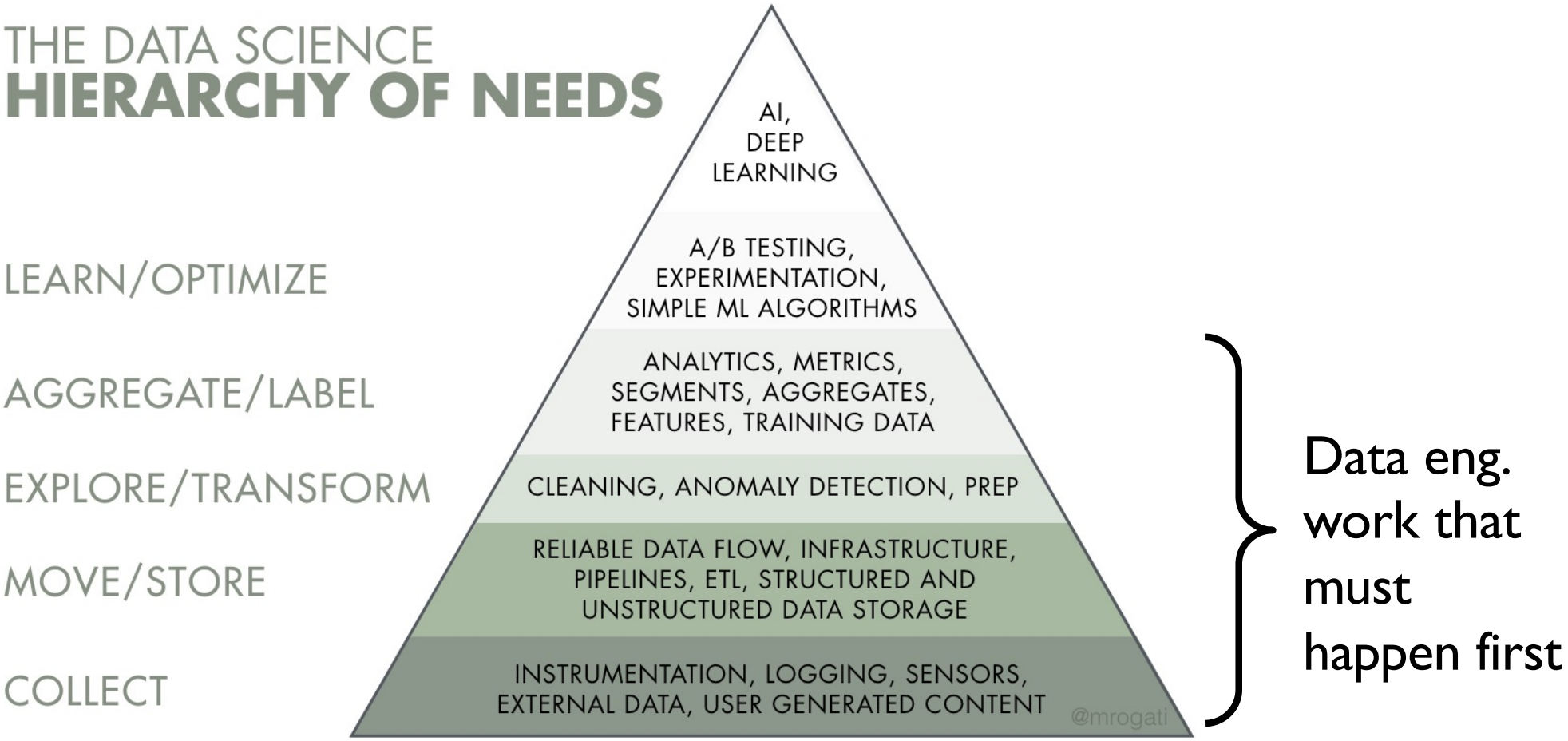
Data Eng Dominates Data Science Projects

Meta

The Llama 3 Herd of Models

"Llama 3 uses a standard, dense Transformer architecture. It does not deviate significantly from Llama and Llama 2 in terms of model architecture; our performance gains are primarily driven by improvements in data quality and diversity as well as by increased training scale."

Data engineer work >> data scientist work



Data eng. work that must happen first

https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

OK, What Does Data Engineering Entail?

ETL and Data Warehouses

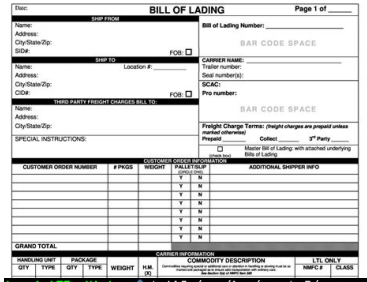
Data lakes

Data Quality

Metadata

Preparation

PDF

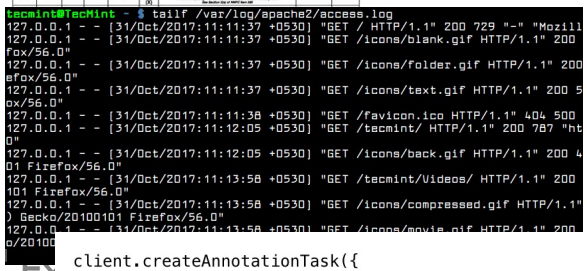


Python Scripts,
Spark/Databricks, etc

Data Preparation

Transform/clean data into useful form

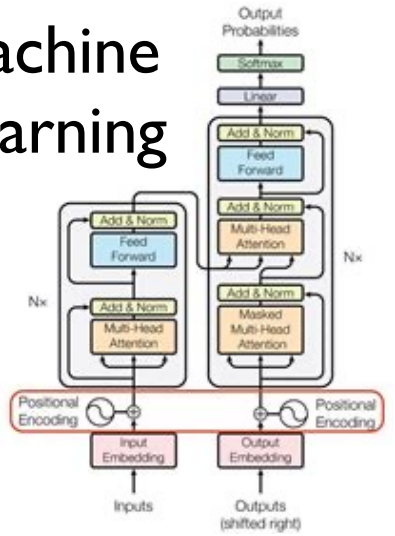
Logs



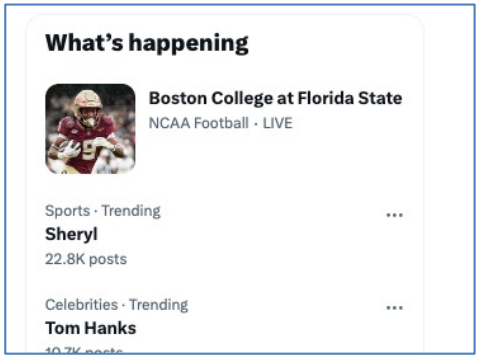

```
client.createAnnotationTask({
  callback_url: 'http://www.example.com/callback',
  instruction: 'Draw a box around each rooftop and pool.',
  attachment: 'http://i.imgur.com/XOJbalC.jpg',
  objects_to_annotate: ['pool', 'rooftop'],
  with_labels: true,
  min_width: 30,
  min_height: 30
}, (err, task) => {
    // do something with task
});
```

JSON

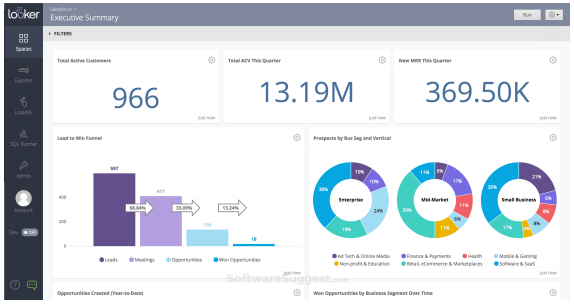
Machine
Learning



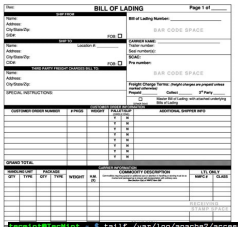
Recommendation

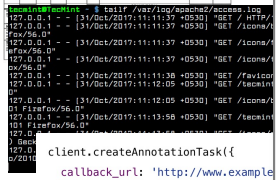


Visualization

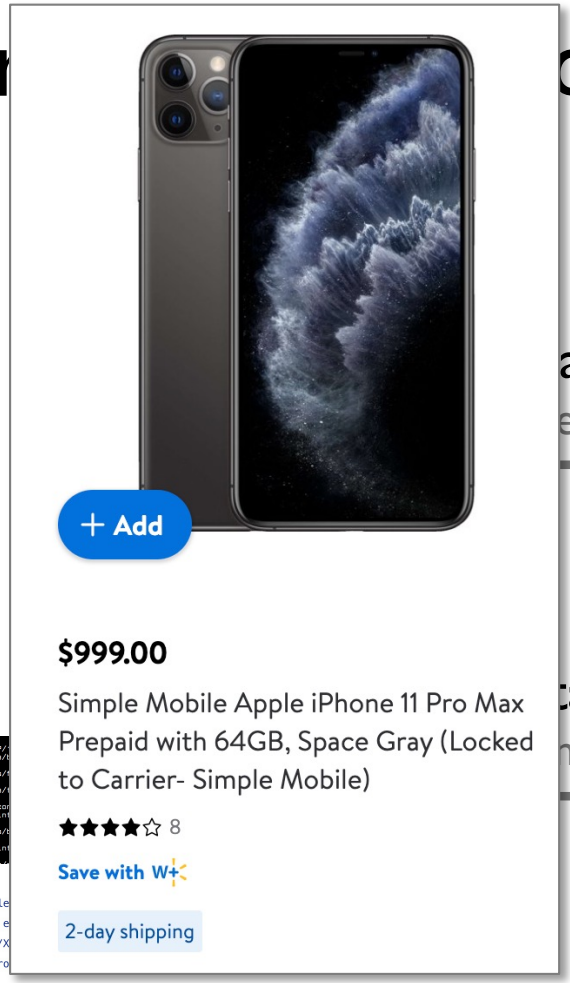


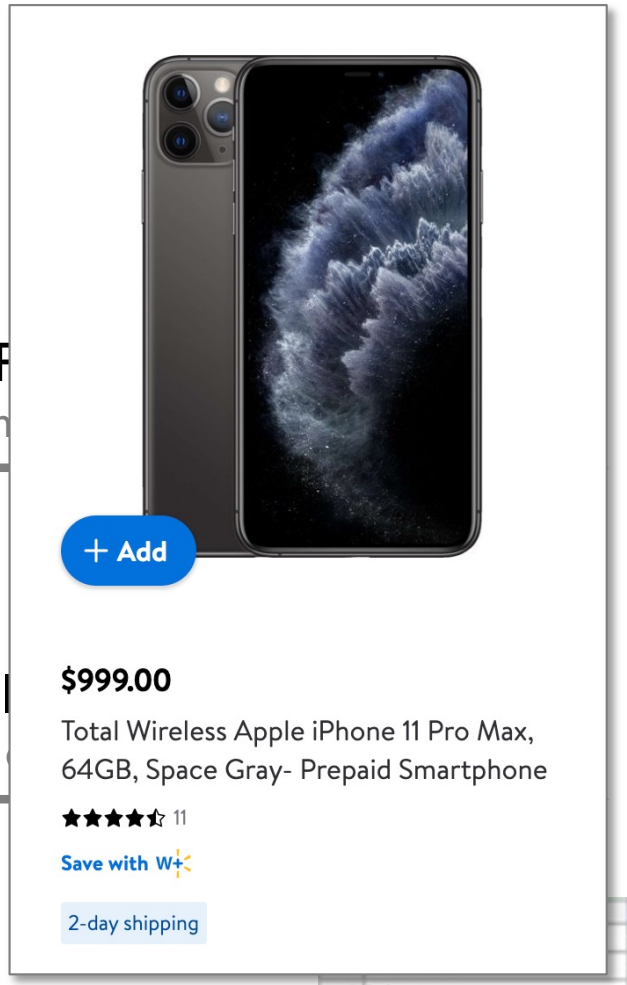
Sources





```
client.createAnnotationTask({
  callback_url: 'http://www.example
  instruction: 'Draw a box around e
  attachment: 'http://i.imgur.com/X
  objects_to_annotate: ['pool', 'ro
  with_labels: true,
  min_width: 30,
  min_height: 30
}, (err, task) => {
  // do something with task
}
```





Use Cases

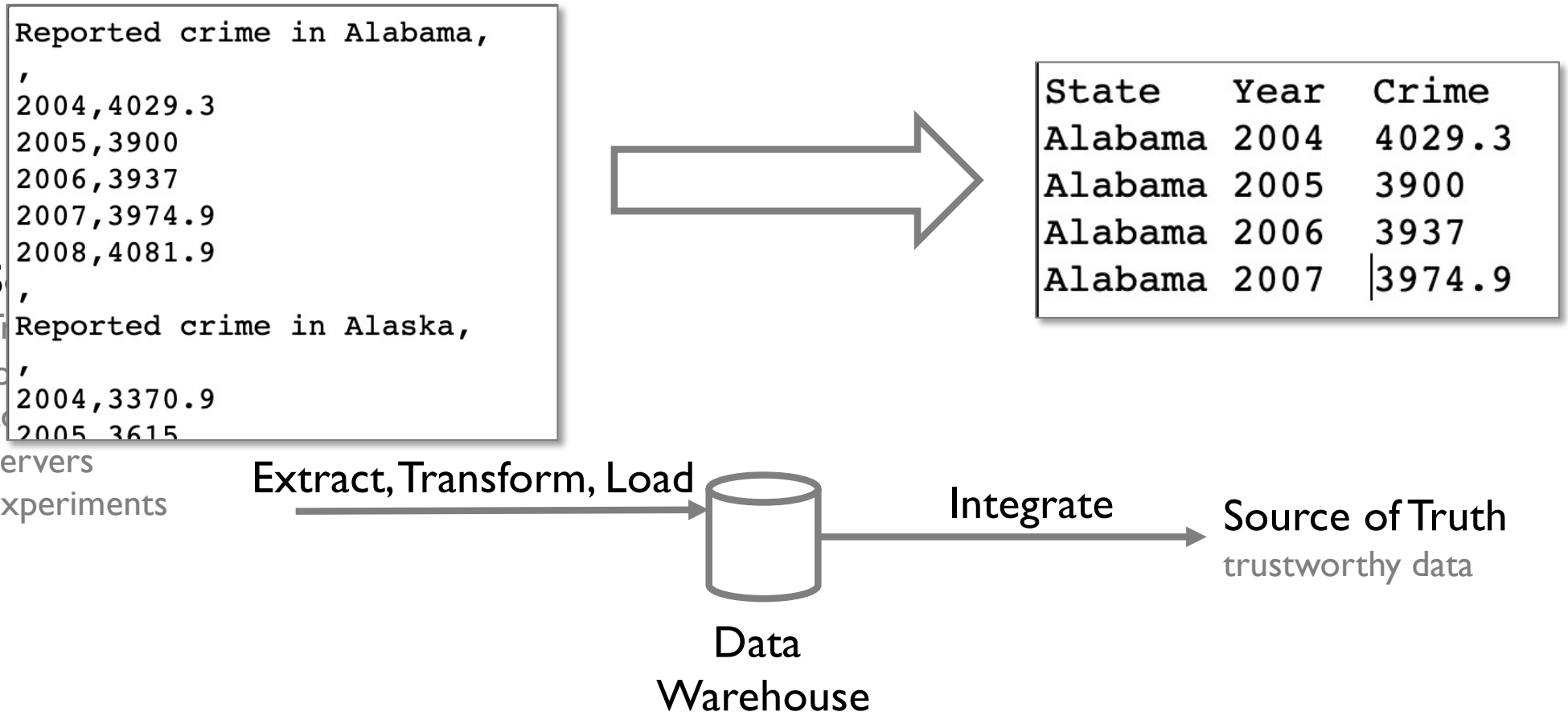
AI, ML, Data science, Apps, Webservices

Source of Truth

trustworthy data

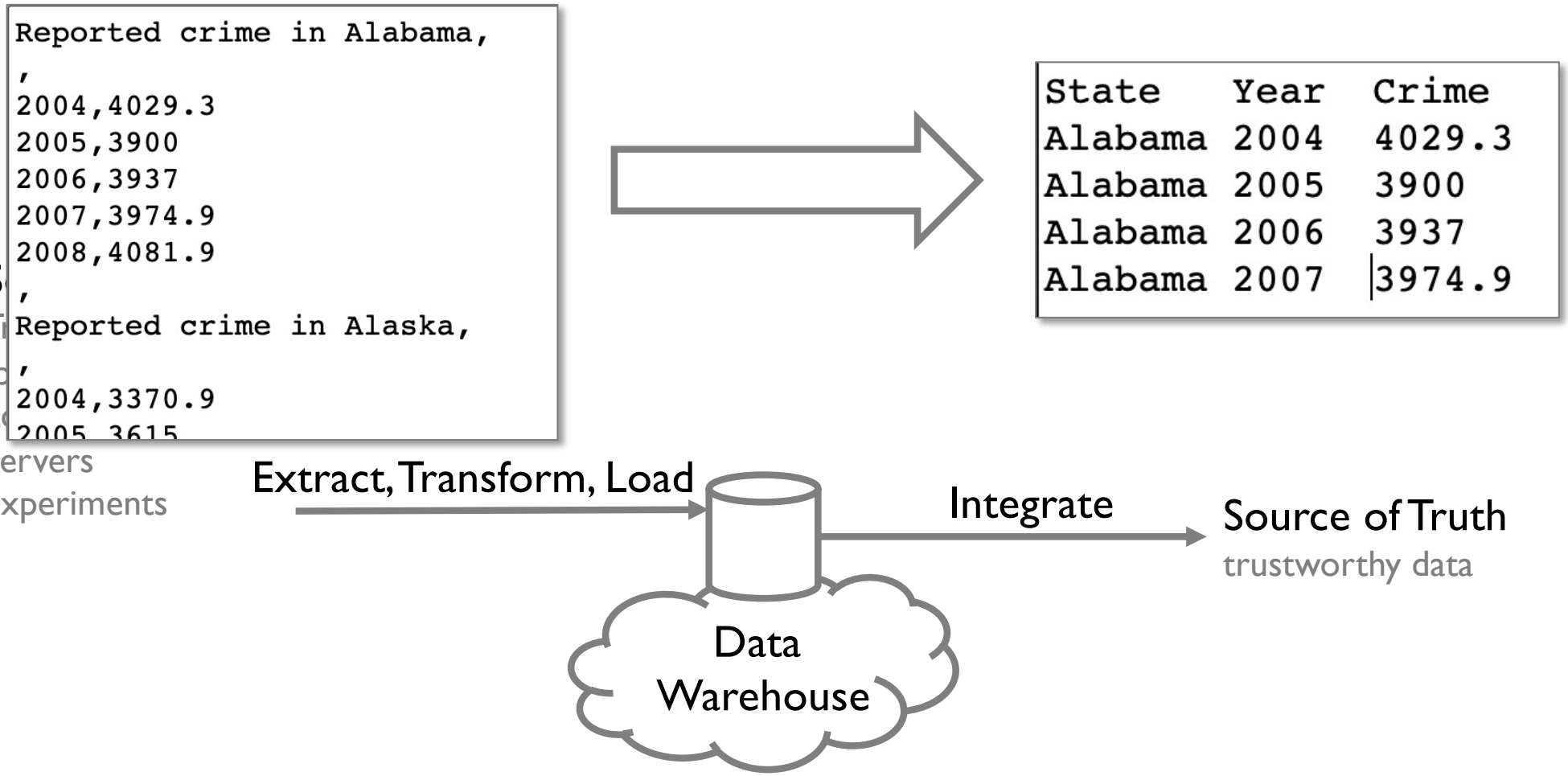
		2004	2005
		4029.3	3900
		3370.9	3615
2	Arizona	5073.3	4827
3	Arkansas	4033.1	4068
4	California	3423.9	3321
5	Colorado	3918.5	4041
6	Connecticut	2684.9	2579
7	Delaware	3283.6	3110
8	District of Columbia	4852.8	4490
9	Florida	4182.5	4013

ETL: Extract Transform Load (1980s)



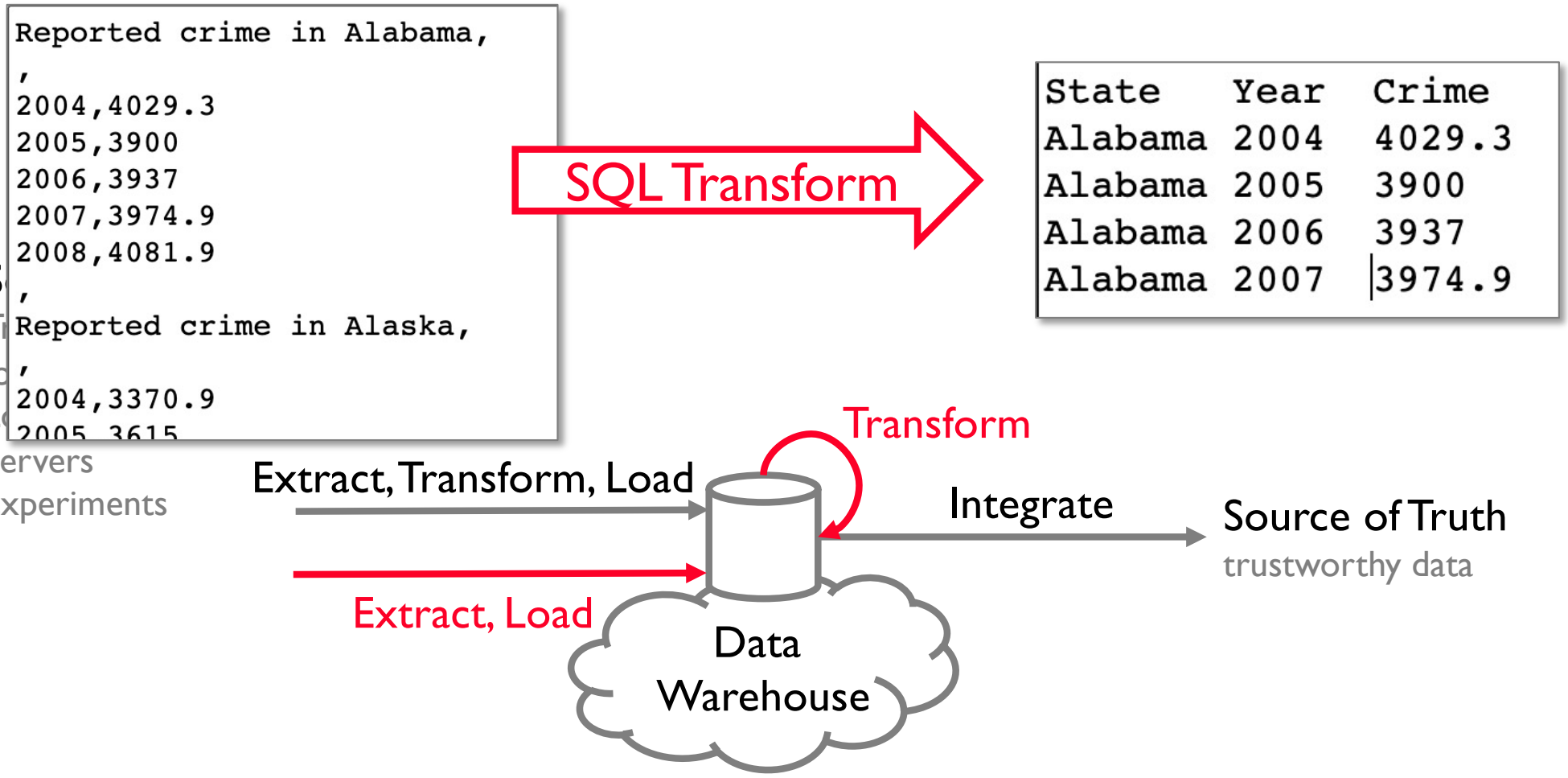
Servers
Experiments

ELT: Extract Load Transform (2010s)



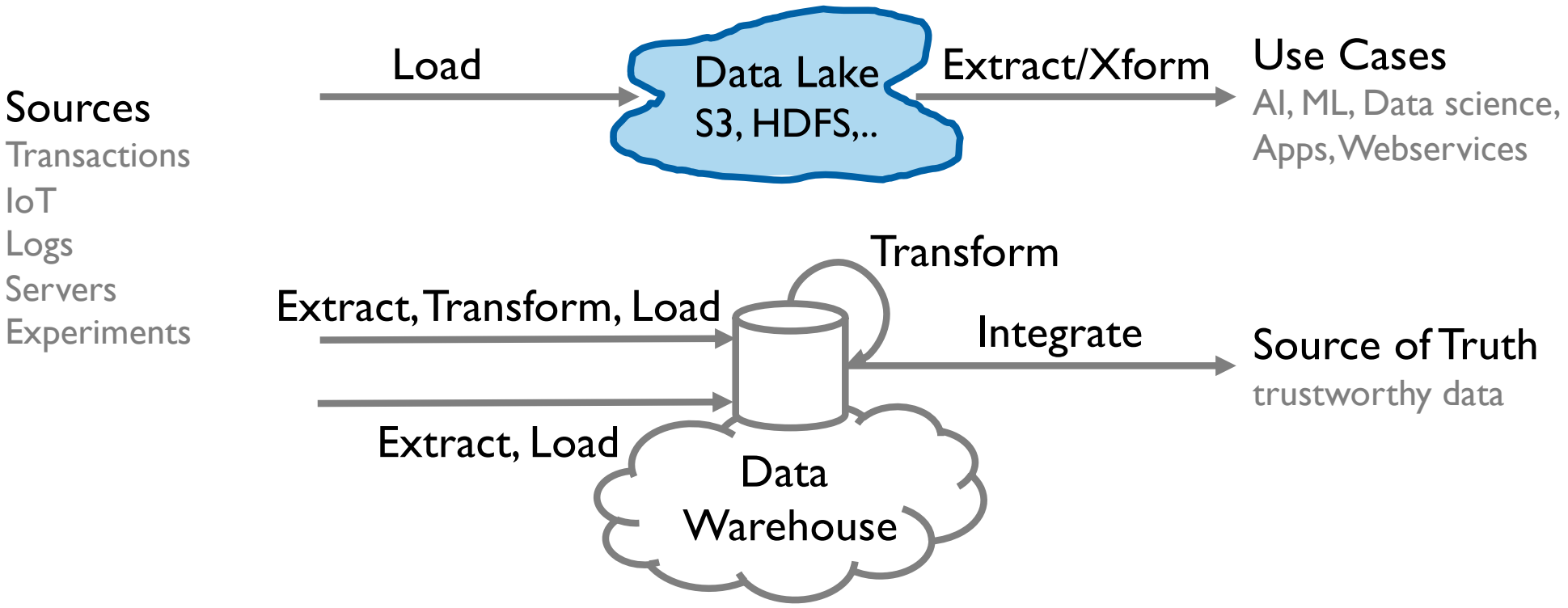
Servers
Experiments

ELT: Extract Load Transform (2010s)

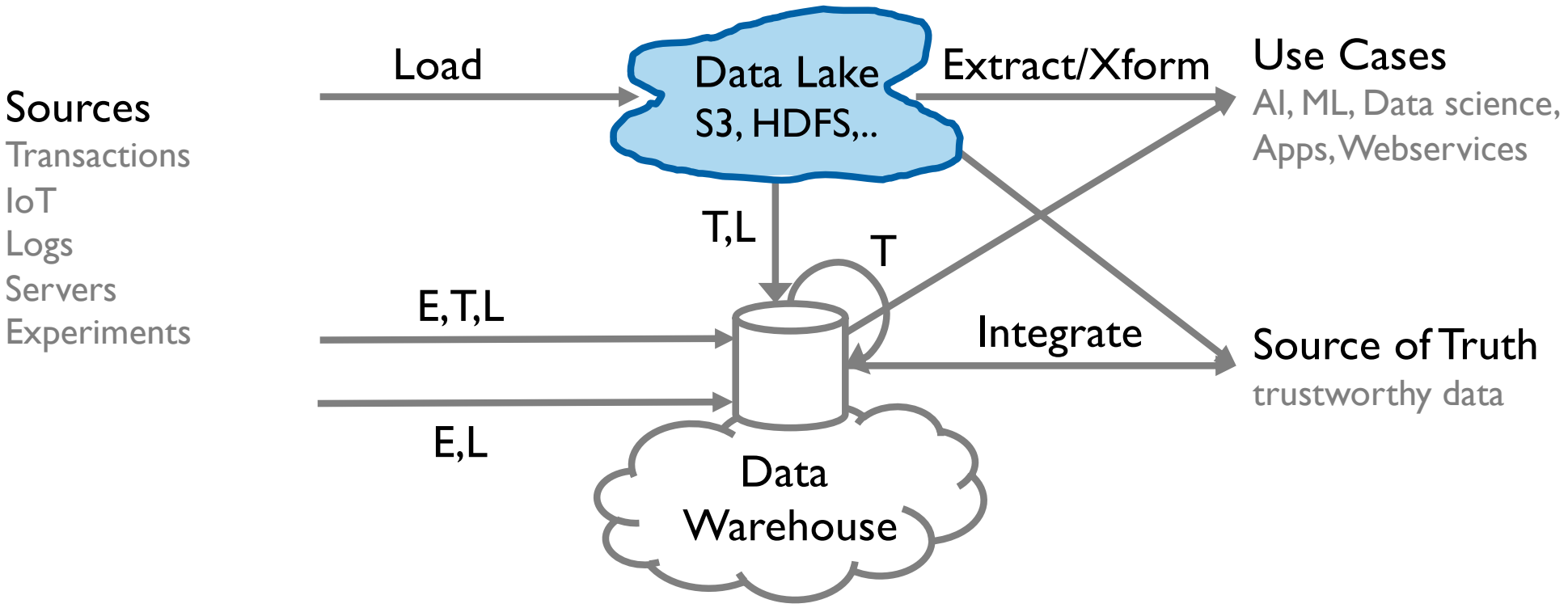


Data Lakes (2000s)

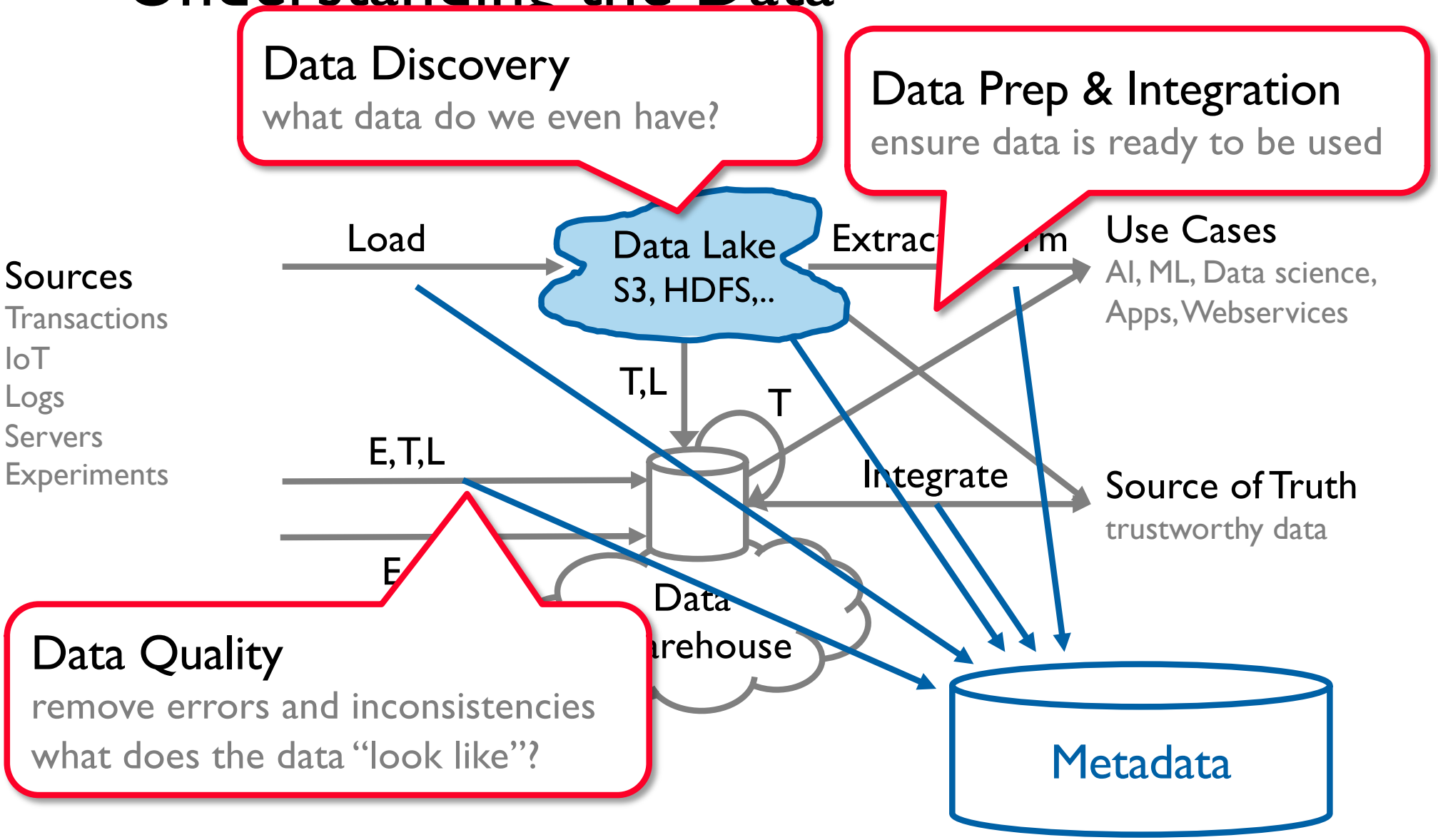
Store as files, transform when needed



Everything Everywhere All At Once (2010s)



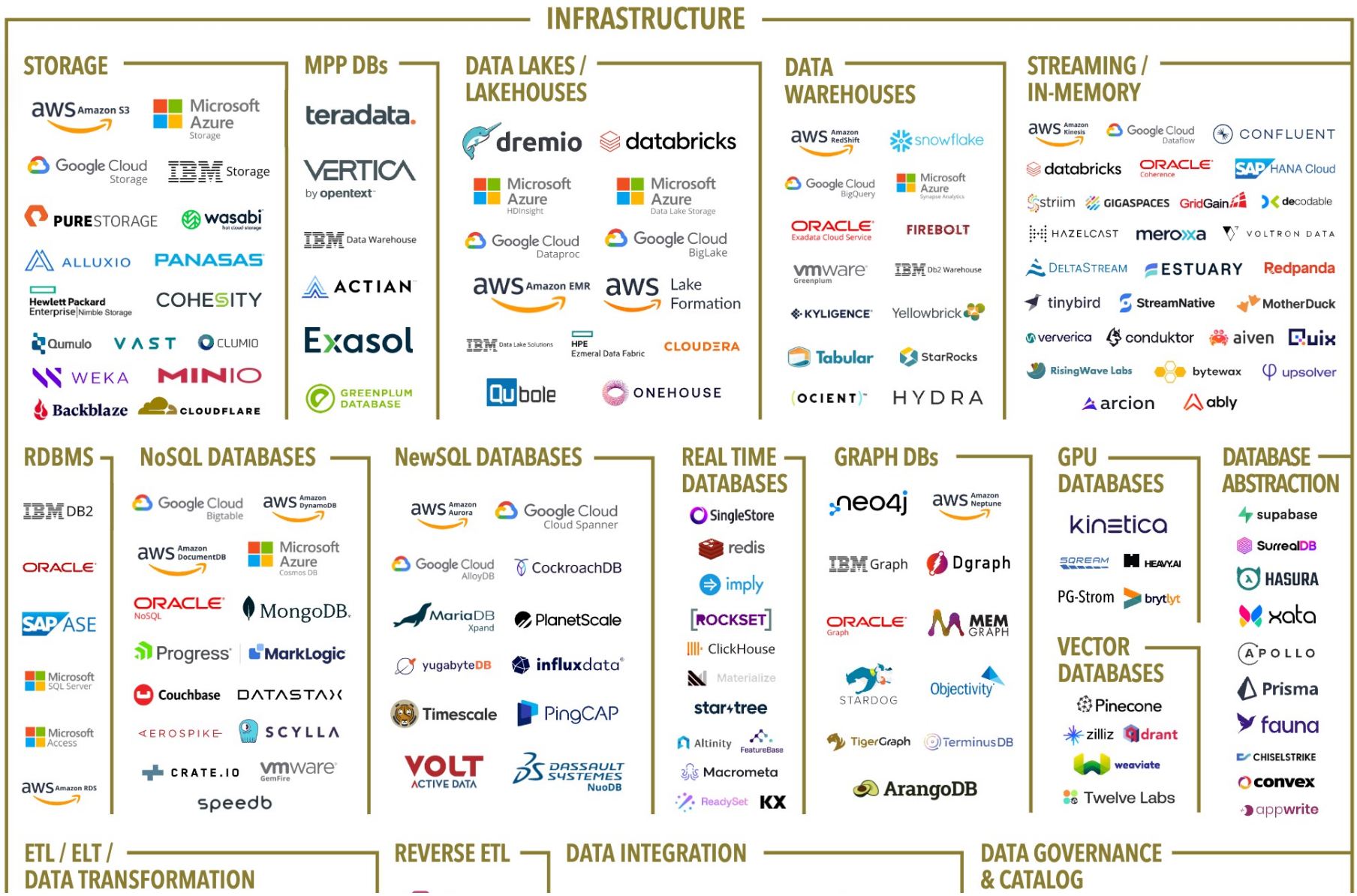
Understanding the Data



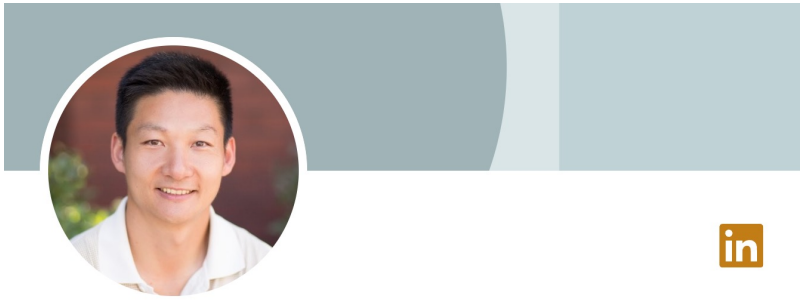

INFRASTRUCTURE
STORAGE
MPP DBs
DATA LAKES / LAKEHOUSES
DATA WAREHOUSES
STREAMING / IN-MEMORY
RDBMS
NoSQL DATABASES
NewSQL DATABASES
REAL TIME DATABASES
GRAPH DB
ETL / ELT / DATA TRANSFORMATION
REVERSE ETL
DATA INTEGRATION
& CATALOG
Pretty much
all products
are derived from
principles in
Relational Data
Management Systems
(this class)
data2020/

Golden Era of Data Systems!

>1000 Database Systems! (from dbdb.io)



Columbia Alumni



Jun Rao
co-founder at Confluent
Confluent . Columbia University

Co-founded $6B stream processing company

John Cieslewicz (He/Him) · 3rd
Data Foundations for SQL, Storage, and Caching at Google
Palo Alto, California, United States · **Contact info**

Heads data infrastructure within google

4111: Intro to Relational Data Management Systems

What's a database?

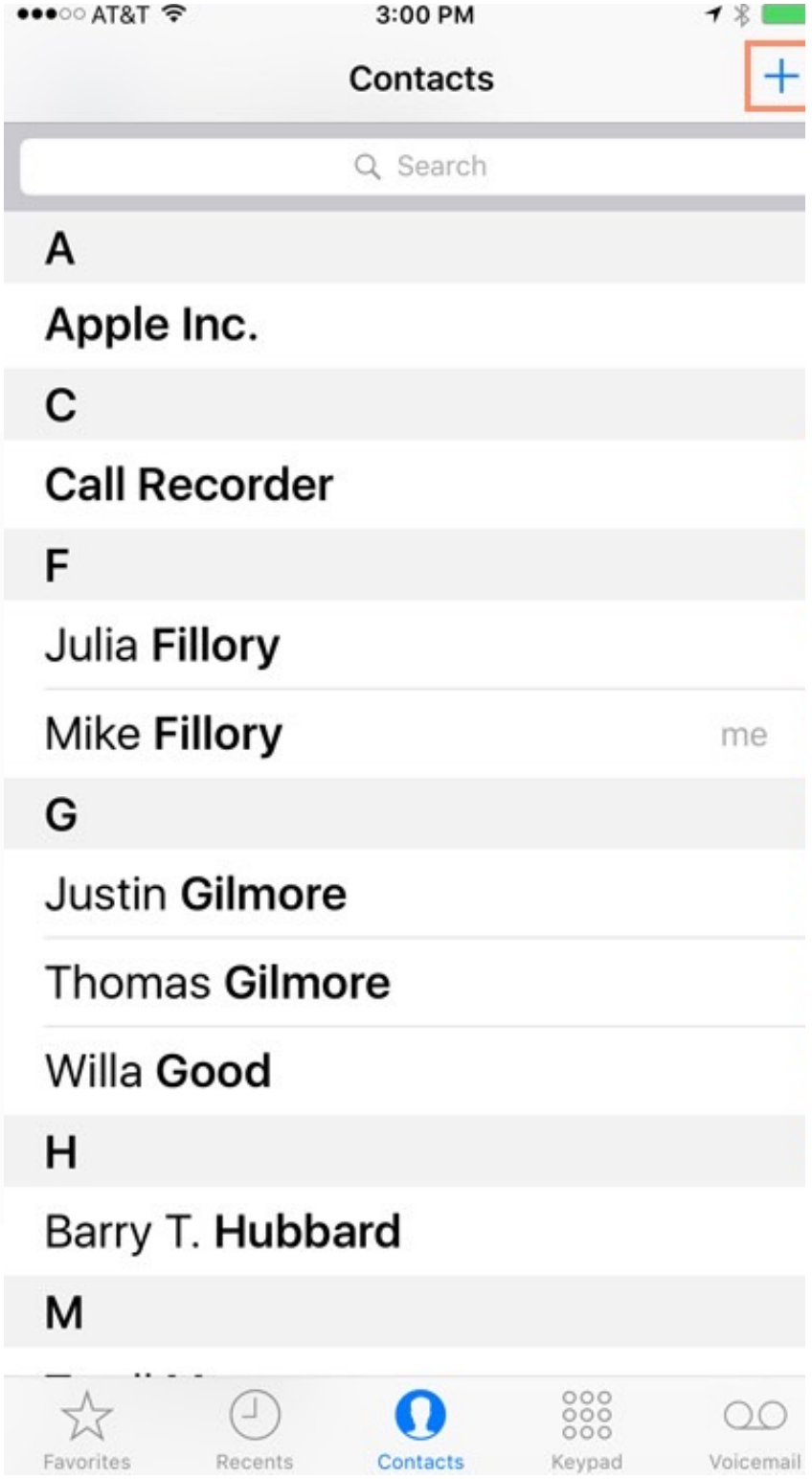
What's a database management system (DBMS)?

What are the core ideas?

What is a Database?

	A	B	C	D	E	F	G	H	I	J	K
1	color	date	slug	title	lshow	link	readings	optional	assigned	ashow	due
2	white	21-Jan	Intro + ER Models			https://githu	Ch 1, 2		HW 0		
3	#e7f8ff	28-Jan	ER Models				Ch 2		<a href="https://github.com/w4111/hw1-s22">H	0	HW 0
4	#e7f8ff	4-Feb	Data Models				Ch 3	optional: <a		0	HW1 Part1
5	#e7f8ff	11-Feb	Data Models + ER->Relational				Ch 3	optional: <a		0	Project 1 Part 1 approval phase
6	#f2f9ed	18-Feb	Relational Algebra				Ch 4		<a href="https://github.com/w4111/project1">Pr	0	Project 1 Part 1 approval phase
7	#f2f9ed	25-Feb	SQL: Basics				Ch 5			0	HW1 Part 2
8	#f2f9ed	4-Mar	SQL: Advanced				Ch 5			0	HW2
9	white	11-Mar	Midterm	one 8x11 page cheat sheet both sides					<a href="https://github.com/w4111/project1/blol	0	
10	white	18-Mar	HOLIDAY							0	Project 1 Part 2
11	#edf3f9	25-Mar	APIs				Ch 6			0	
12	#edf3f9	1-Apr	Data Quality	Normalization and data errors			Ch 19		<a href="https://github.com/w4111/hw4-s22">H	0	HW3
13	#ddf9ff	8-Apr	Physical Design				Ch 8		<a href="https://github.com/w4111/project2_s22	0	
14	#ddf9ff	15-Apr	Query Processing				Ch 12			0	Project 1 Part 3
15	#ddf9ff	22-Apr	Transactions				Ch 16, 18			0	
16	white	29-Apr	Data Pipelines							0	HW 4
17	white	13-May	Exam 2 (Cum	one 8x11 page cheat sheet both sides						0	Project 2
18											
19											
20											

What is a Database?



What is a Database?

```
2012-01-04 00:01:23,180 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-2281137920769
010
2012-01-04 00:01:23,184 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32981,
 cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-228113
2012-01-04 00:01:23,185 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-2
2012-01-04 00:01:23,291 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_37660314352523
10
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32982,
cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_37660314
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_37
2012-01-04 00:01:23,324 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-8044922265890
010
2012-01-04 00:01:23,326 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32983,
liID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-80449222
2012-01-04 00:01:23,327 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-8
2012-01-04 00:01:23,409 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-9657937572621
10
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32984,
, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-96579
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-9
2012-01-04 00:01:23,433 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:50010,
 cliID: DFSClient_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_-96579
2012-01-04 00:01:23,494 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_54159109576590
10
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32987,
, cliID: DFSClient_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid: blk_54159
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_54
2012-01-04 00:01:23,523 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-5517241460358
```

What is a Database?

What is a Database?

Lots of

Structured data

Database Management System (DBMS)

A system to store, manage and access databases

Database Management System (DBMS)

System to safely and reliably store lots of persistent structured data and is convenient for multiple users to efficiently access and modify.

Is a script a DBMS?

Javascript/Python Script

- Data stored in variables (RAM)
- Very fast access
- Data structures (lists, dicts, tuples)

Is Excel a DBMS?

Microsoft office security

Visually access/modify/compute over data cells

Click save to store persistently

Is the file system a DBMS?

Manages files that are persistently stored on disk

Open/read/seek/write access to files

Access via file names

Access control via permissions

Is the file system a DBMS?

You and a friend edit the same text file

Save at the same time

What happens?

1. Your changes survive
2. Friend's changes survive
3. Both changes survive
4. No changes survive
5. ¯_(ツ)_/¯

Is the file system a DBMS?

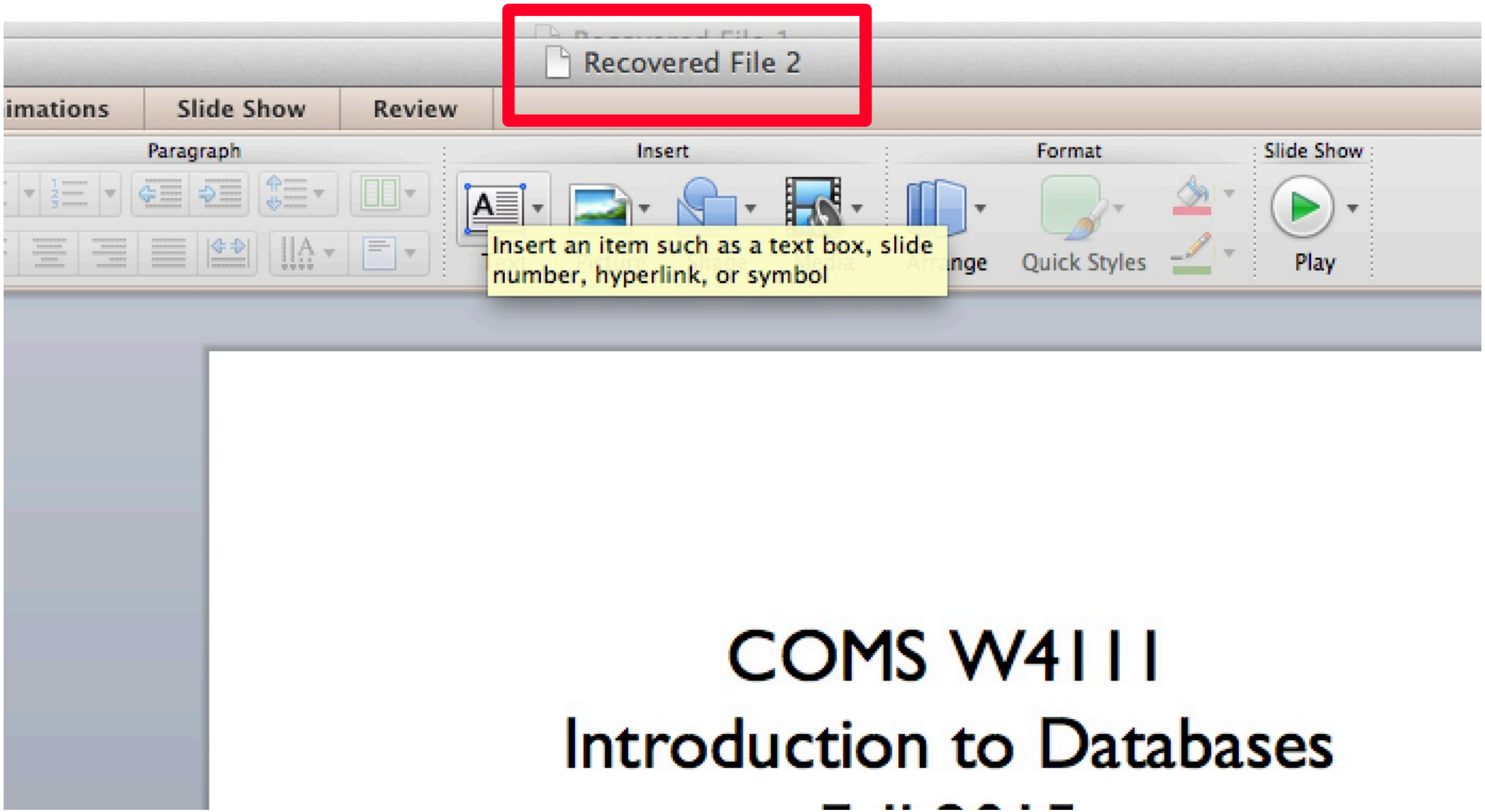
You edit a text file

Computer crashes

What happens?

1. All changes survive
2. No changes survive
3. Changes from last save survive
4. ¯_(ツ)_/¯

Is the file system a DBMS?

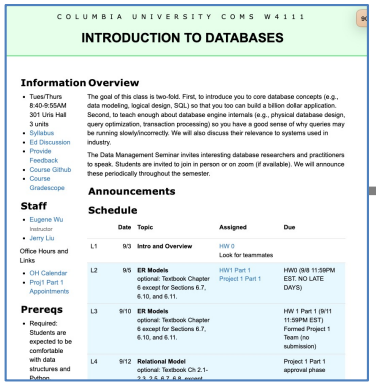


Who… would ever do this?

Real $1B+ Companies…

Store extracted data in a file

Every change → rewrite the file


COLUMBIA UNIVERSITY COMS W4111

INTRODUCTION TO DATABASES

Information Overview

- Tues/Thurs 8:40-9:55AM 301 Uris Hall 3 units
- Syllabus
- Ed Discussion
- Provide Feedback
- Course Github
- Course Gradescope

The goal of this class is two-fold. First, to introduce you to core database concepts (e.g., data modeling, logical design, SQL) so that you too can build a billion dollar application. Second, to teach enough about database engine internals (e.g., physical database design, query optimization, transaction processing) so you have a good sense of why queries may be running slowly/incorrectly. We will also discuss their relevance to systems used in industry.

The Data Management Seminar invites interesting database researchers and practitioners to speak. Students are invited to join in person or on zoom (if available). We will announce these periodically throughout the semester.

Announcements

Schedule

	Date	Topic	Assigned	Due
L1	9/3	Intro and Overview	HW 0 Lock for teammates	
L2	9/5	ER Models optional: Textbook Chapter 6 except for Sections 6.7, 6.10, and 6.11.	HW1 Part 1 Project 1 Part 1	HW0 (9/8 11:59PM EST. NO LATE DAYS)
L3	9/10	ER Models optional: Textbook Chapter 6 except for Sections 6.7, 6.10, and 6.11.		HW 1 Part 1 (9/11 11:59PM EST) Formed Project 1 Team (no submission)
L4	9/12	Relational Model optional: Textbook Ch 2.1-2.3, 2.6, 2.7, 6.8, excerpt		Project 1 Part 1 approval phase

Staff

- Eugene Wu Instructor
- Jerry Liu

Office Hours and Links

- OH Calendar
- Proj1 Part 1 Appointments

Prereqs

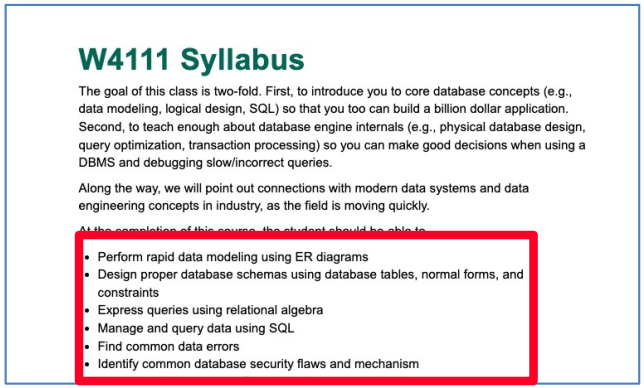
- Required: Students are expected to be comfortable with data structures and Python

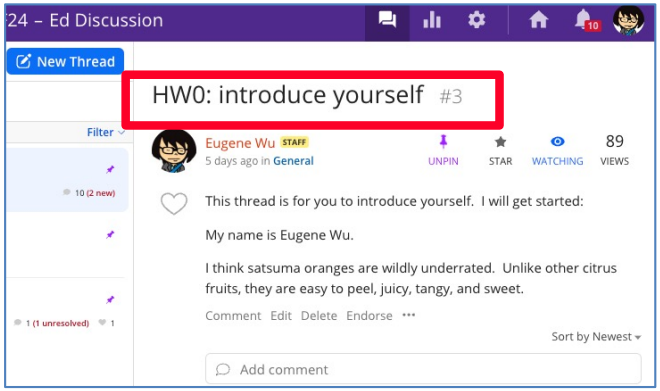
Text extraction and transform tasks

```
[{                          [{
  obj: region                 obj: authorlist
  box: 36,34,100,120          box: 40,40,100,20
  text: “Private F..”         text: “Eugene W..”
 }, { .. }, ..]             }, { .. }, ..]
```

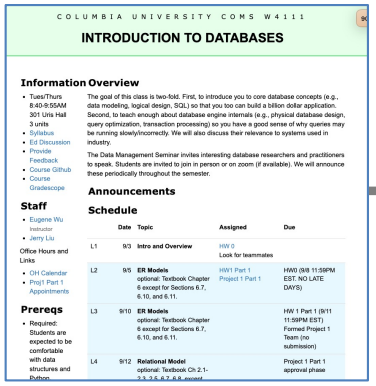
Browser

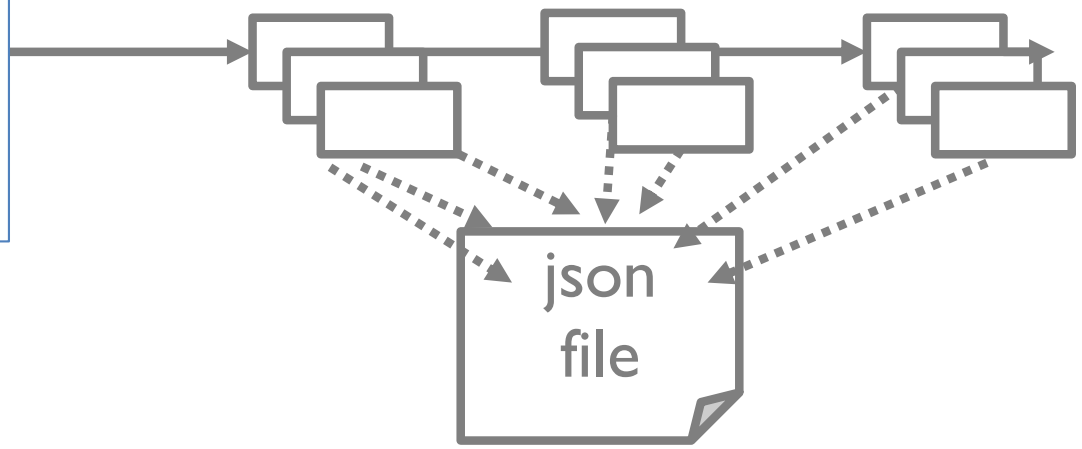
New Tab





Annotation tasks





Want *Guarantees* from DBMS

You want to write a hot new app on a DBMS.
What do you *not* want to worry about?

Failures disk, machine, human, corruption, deity

Lots of users concurrency, scaling, responsiveness

Ad-hoc data access arbitrary queries

Data formats csv? tsv? custom format?

Database Management System (DBMS)

System to safely and reliably store lots of persistent structured data and is convenient for multiple users to efficiently access and modify.

Database Management System (DBMS)

Safe	Consistent and correct data after failures
Reliable	99.99+% Uptime
Lots	>>RAM (petabytes, exabytes, zetabytes..)
Persistent	Lives longer than DBMS application
Convenient	Physical Independence. Declarative.
Multiple Users	Concurrent access. Access control.
Efficient	*Fast*: 100k+ queries / sec

You *Want* These Properties…

For anything important!

$$: Banking, Commerce, Insurance, Stocks, …

Social: messages, videos, likes, friends, voting,…

Science: surveys, drug trials, experiments, sensors, …

Industry: fleets, supply chain, jobs, warehouse,..

Companies: hiring, insurance, payroll, security, …

DB Encompasses Most of CS

OS	DBMS directly manages hardware
Languages	SQL is a domain specific language
Theory	Algorithms, models, NP-complete
AI/ML	Knowledge Discovery, KDD,NL2SQL
Logic	Relational Algebra = 1st order logic

Scalable Computer Science

Key Concepts

Data Independence

Declarative Languages

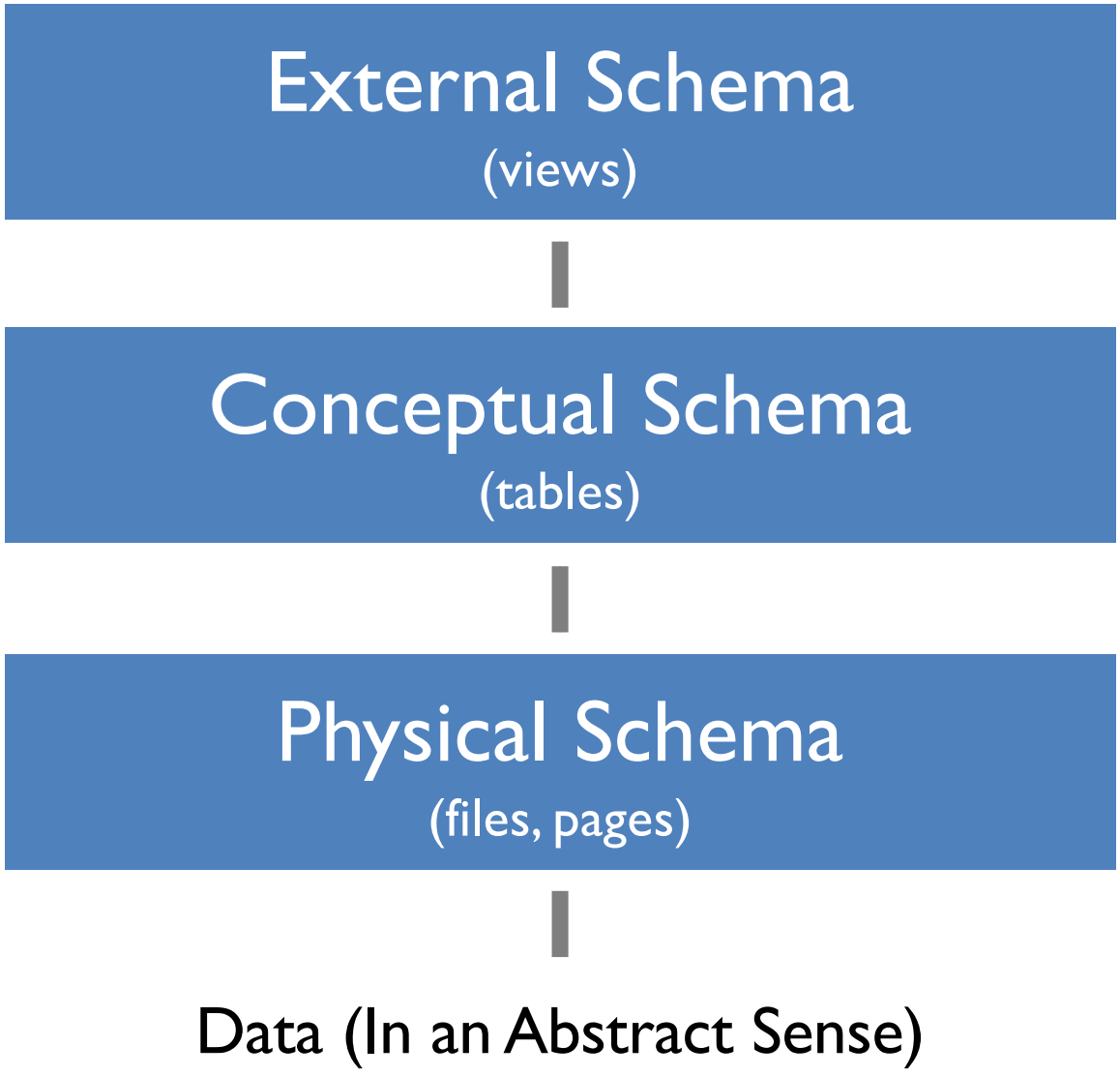
Serve to insulate application programmers from the system implementation

Data Independence

External Schema
How users/app see data

Conceptual Schema
The logical structure

Physical Schema
How data are stored



Example App: Ruber

Users(uid int, name str, age int)

Drivers(did int, name str)

Rides(uid int, did int, distance float, drive_time float)

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0,Eugene,17
1,Luis,20
2,Ken,30

CSV File

What is the number of adults?

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0,Eugene,17
1,Luis,20
2,Ken,30

CSV File

```
n = 0
for line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

0 Eugene 17
1 Luis 20
2 Ken 30

TSV File

```
n = 0
for line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence

UID	Name	Age
0	Eugene	17
1	Luis	20
2	Ken	30

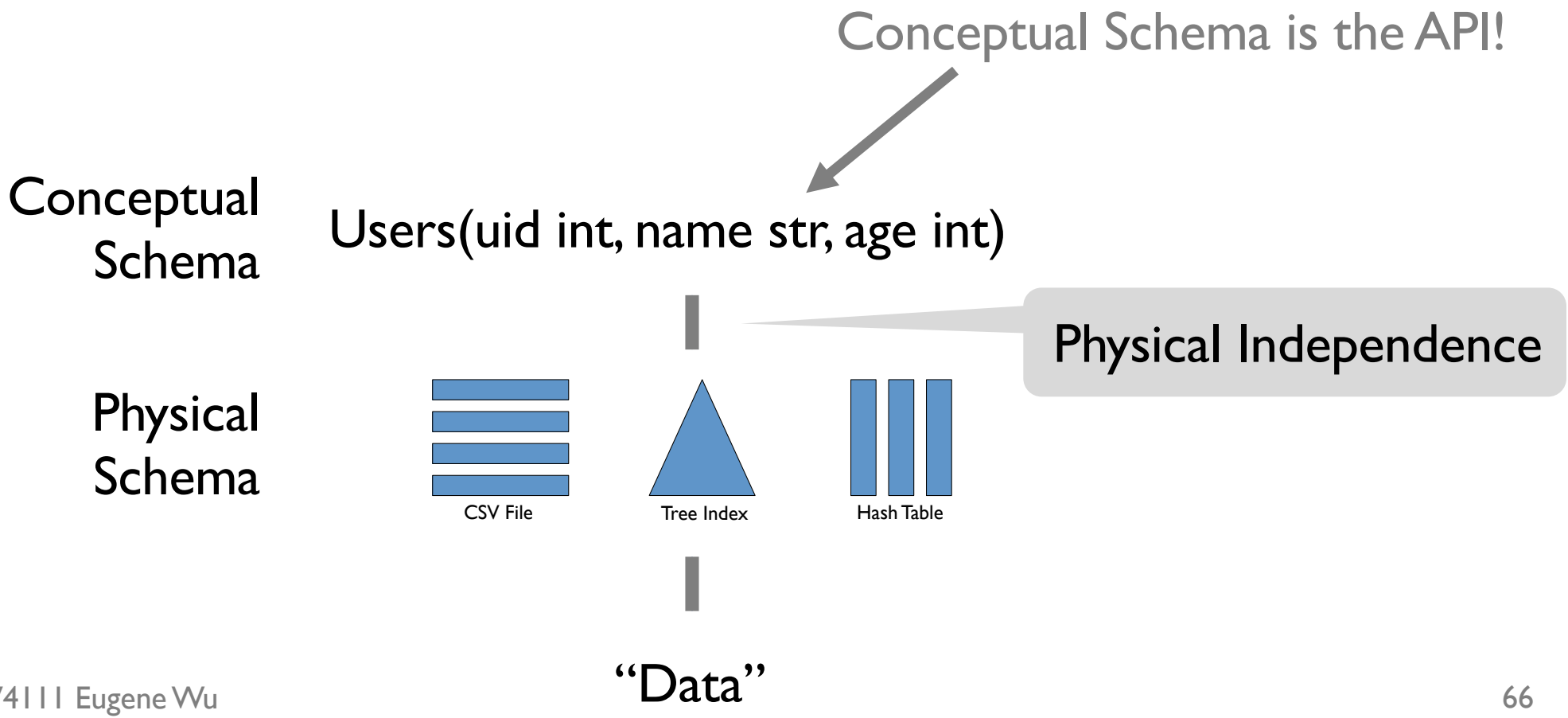
0,1,2
Eugene,Luis,Ken
17,20,30

Columnar File

```
n = 0
for line in csv_file:
  attributes = line.split(“,”)
  if attributes[2] >= 18:
    n += 1
```

Data Independence

```
print(f“Welcome back Mr. {name.split()[-1]}”)
> “Welcome back Mr. Bond”
```



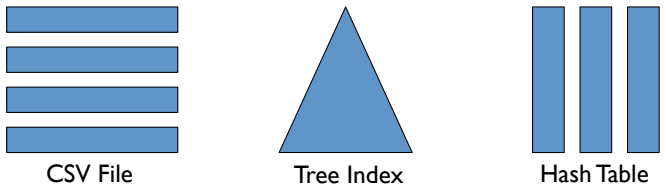
Data Independence

~~`print(f"Welcome back Mr. {name.split()[-1]}")`~~

Conceptual Schema

Users(uid int, fname str, lname str, age int)

Physical Schema



Physical Independence

"Data"

Data Independence

```
print(f“Welcome back Mr. {name.split()[-1]}”)
> “Welcome back Mr. Bond”
```

External Schema

Users(uid int, name str, age int)

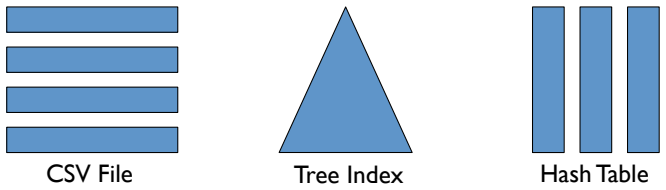
Logical Independence

Conceptual Schema

Users(uid int, fname str, lname str, age int)

Physical Independence

Physical Schema



“Data”

Data Independence

One of most important properties of a DBMS

Protection from changes in logical structure of data — Logical Independence

Protection from changes in physical structure of data — Physical Independence

Declarative Interface

Mechanism that enables data independence

Insulates programmer from physical schema

Rather than a list of functions,

the API is a *query language*

(a domain specific language)

Declarative Interface

What you want,

"Make me a sandwich"

Buy from pb&j store
Make BLT
½ Tuna
Veggie

not how to do it.

"Take two slices of wheat bread out of the 2nd shelf, put them next to each other…"

What if on 1st shelf?
Out of wheat bread?
No counter space?

Declarative Interface

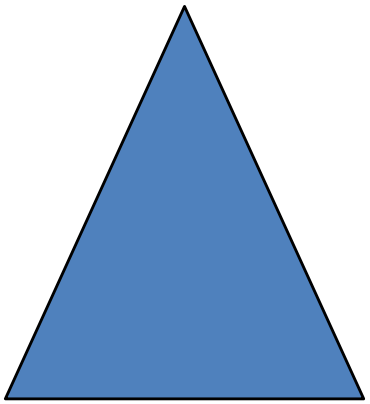
"I want all highly rated fast drivers"

Declarative Interface

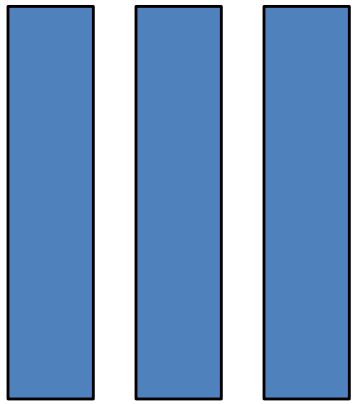
SELECT name FROM users WHERE rating > 8

DBMS

CSV File



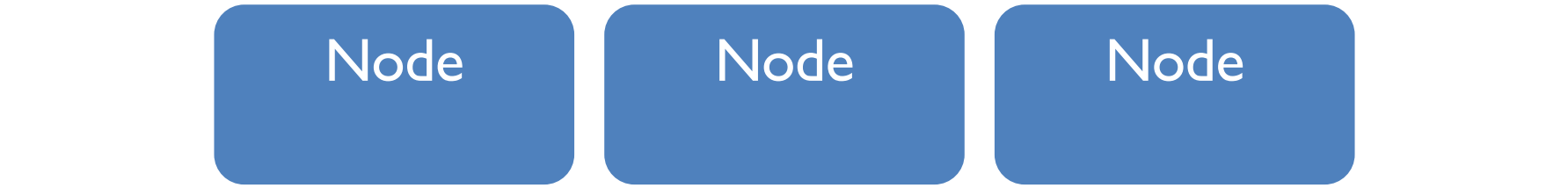
Tree Index



Hash Table

Declarative Interface

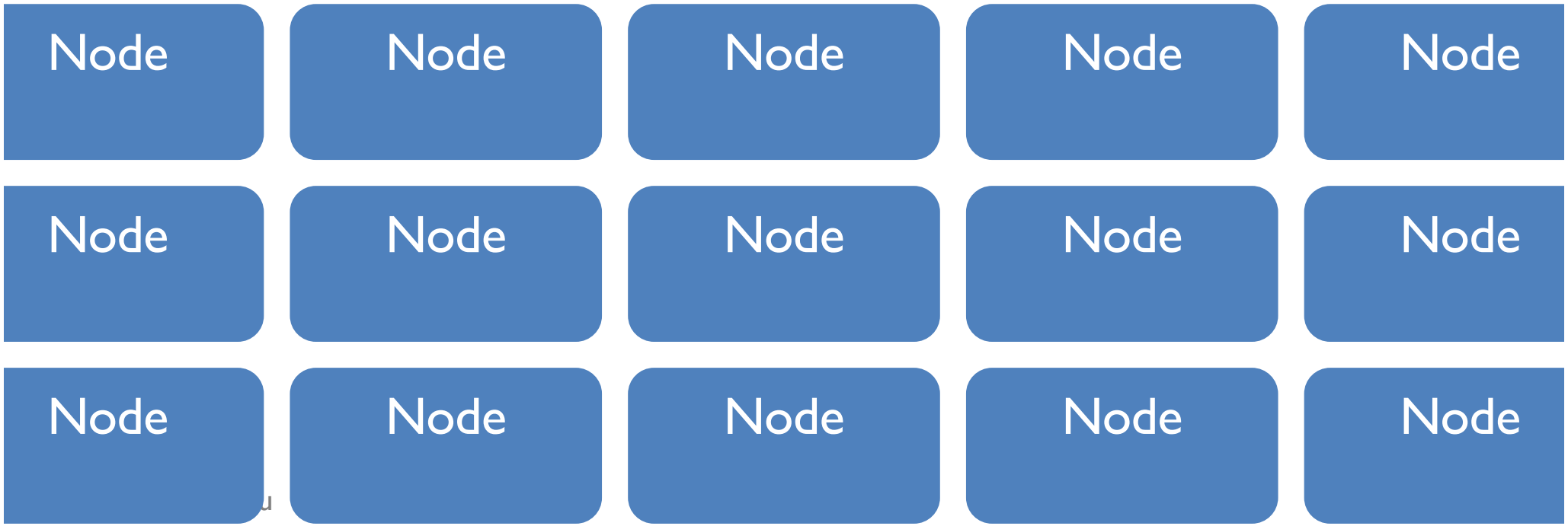
SELECT name FROM users WHERE rating > 8

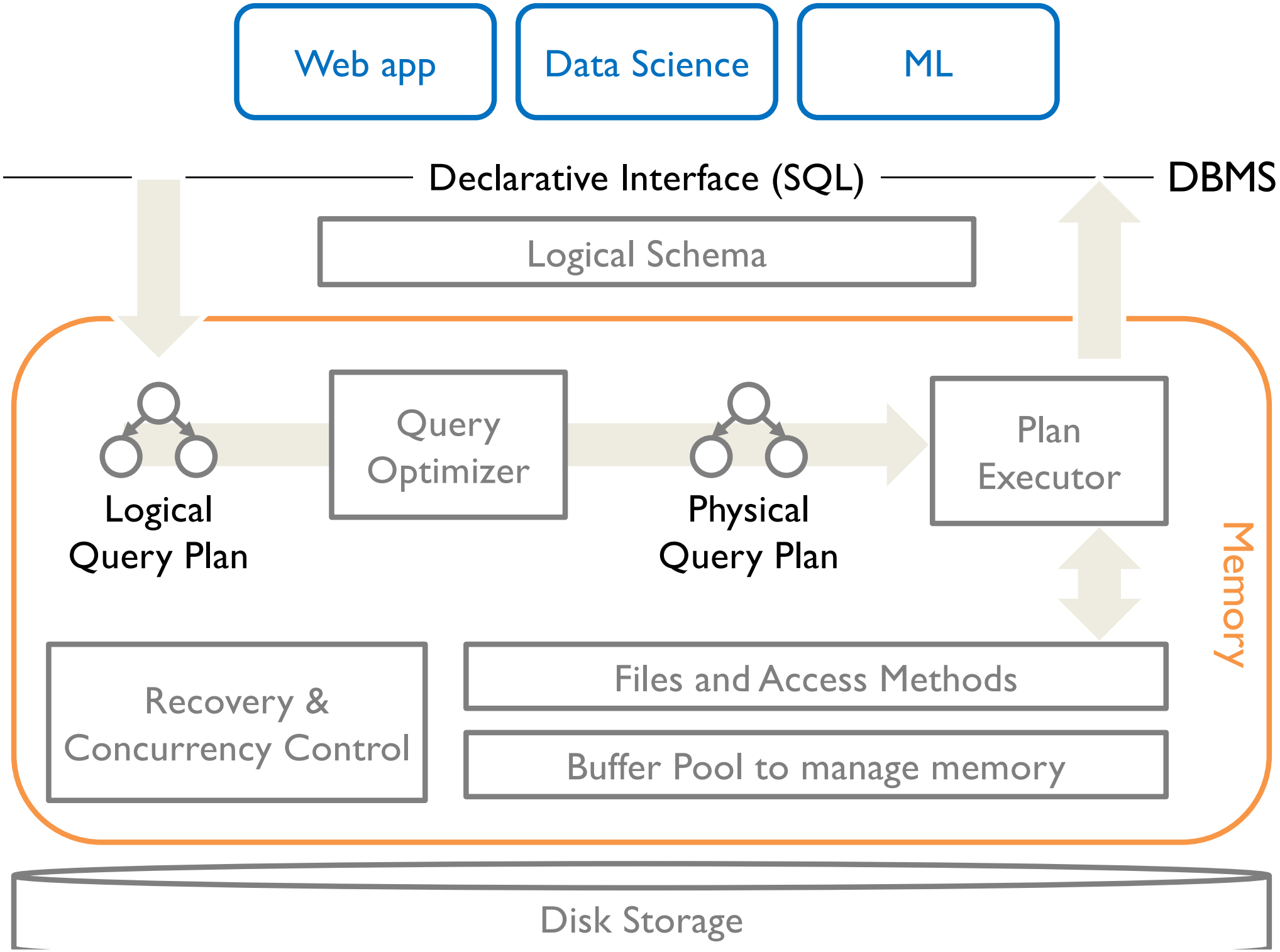


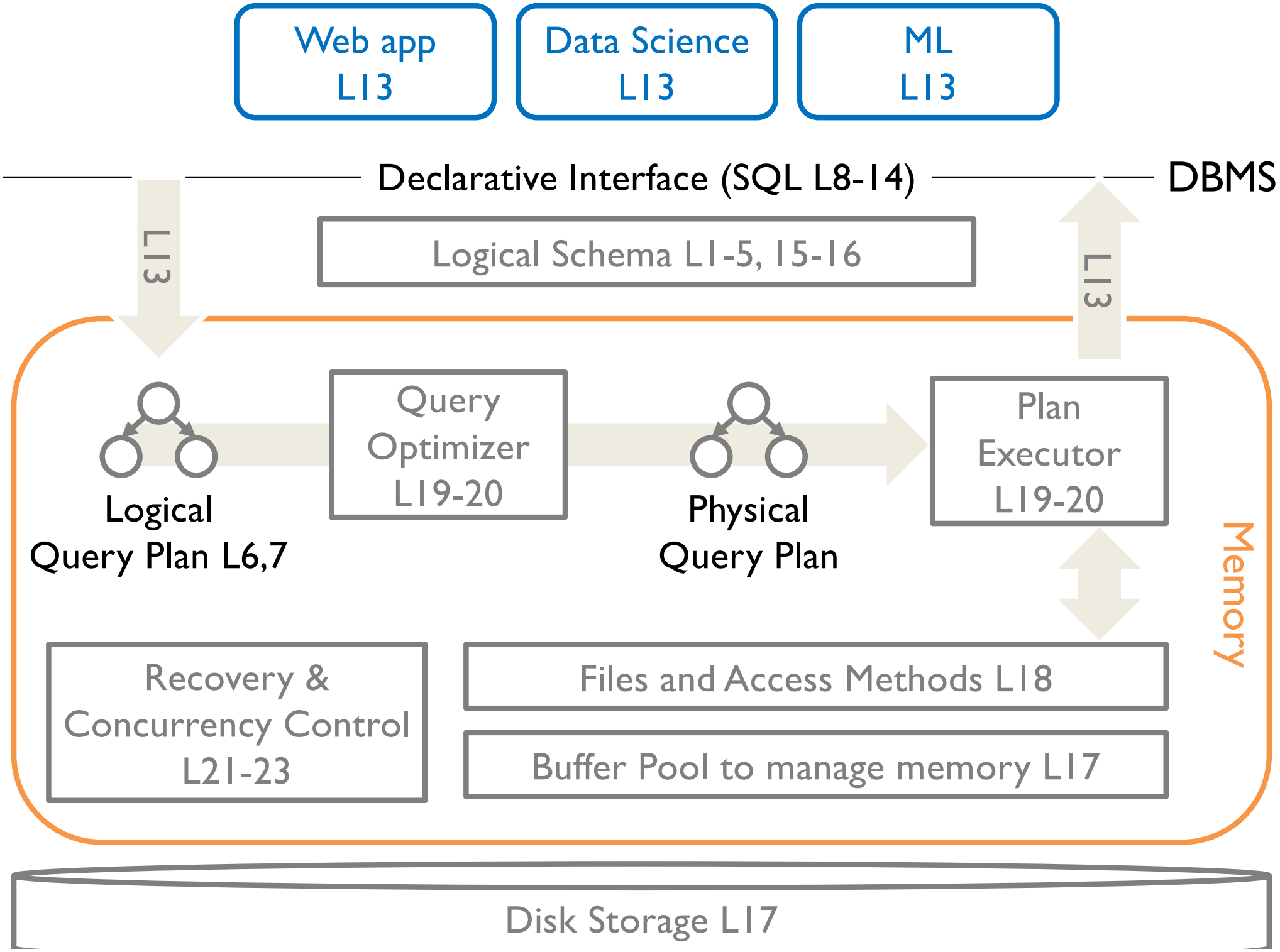
Declarative Interface

SELECT name FROM users WHERE rating > 8

DBMS




Web app
Data Science
ML
Declarative Interface (SQL)
DBMS
Logical Schema
Query Optimizer
Logical Query Plan
Physical Query Plan
Plan Executor
Memory
Recovery & Concurrency Control
Files and Access Methods
Buffer Pool to manage memory
Disk Storage


Web app
L13
Data Science
L13
ML
L13
Declarative Interface (SQL L8-14)
DBMS
L13
Logical Schema L1-5, 15-16
L13
Query
Optimizer
L19-20
Plan
Executor
L19-20
Logical
Query Plan L6,7
Physical
Query Plan
Memory
Recovery &
Concurrency Control
L21-23
Files and Access Methods L18
Buffer Pool to manage memory L17
Disk Storage L17

A bit about the class

w4111.github.io

COLUMBIA UNIVERSITY COMS W4111

INTRODUCTION TO DATABASES

Information

- Tues/Thurs
  8:40-9:55AM
  301 Uris Hall
  3 units
- Syllabus
- Ed Discussion
- Provide Feedback
- Course Github
- Course Gradescope

Staff

- Eugene Wu Instructor
- Jerry Liu

Office Hours and Links

- OH Calendar
- Proj1 Part 1 Appointments

Prereqs

- Required: Students are expected to be comfortable with data structures and Python.
- Required: COMS W3134, COMS W3137, or COMS

Overview

The goal of this class is two-fold. First, to introduce you to core database concepts (e.g., data modeling, logical design, SQL) so that you too can build a billion dollar application. Second, to teach enough about database engine internals (e.g., physical database design, query optimization, transaction processing) so you have a good sense of why queries may be running slowly/incorrectly. We will also discuss their relevance to systems used in industry.

The Data Management Seminar invites interesting database researchers and practitioners to speak. Students are invited to join in person or on zoom (if available). We will announce these periodically throughout the semester.

Announcements

Schedule

	Date	Topic	Assigned	Due
L1	9/3	**Intro and Overview**	HW 0 Look for teammates	
L2	9/5	**ER Models** optional: Textbook Chapter 6 except for Sections 6.7, 6.10, and 6.11.	HW1 Part 1 Project 1 Part 1	HW0 (9/8 11:59PM EST. NO LATE DAYS)
L3	9/10	**ER Models** optional: Textbook Chapter 6 except for Sections 6.7, 6.10, and 6.11.		HW 1 Part 1 (9/11 11:59PM EST) Formed Project 1 Team (no submission)
L4	9/12	**Relational Model** optional: Textbook Ch 2.1-2.3, 2.5, 6.7, 6.8, except 6.7.2		Project 1 Part 1 approval phase

Next Up

HW0 is out.

https://github.com/w4111/hw0

Due by 9/8 11:59PM.

Late Submissions = -3% on final grade

Class Information

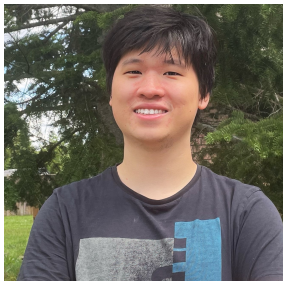
Tu/Th 8:40-9:55AM in 301 Uris

Prereqs:

- COMS W3134 - *Data Structures in Java* or
- COMS W3137 - *Data Structures and Algorithms*

  (equivalent courses taken elsewhere are acceptable as well)
- Fluency in Python

Your TAs

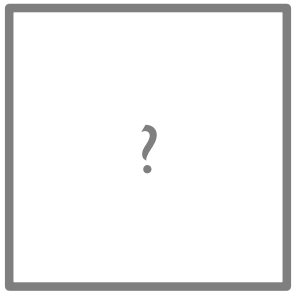


Jerry Huang. PhD. Head TA

I like watching anime and playing chess. One of my favorite anime is Steins;Gate and my favorite chess opening is the Marshall Attack variation in Ruy Lopez opening.

Jordyn Jaffe

My fun fact is that I also work for Columbia Business School.



???

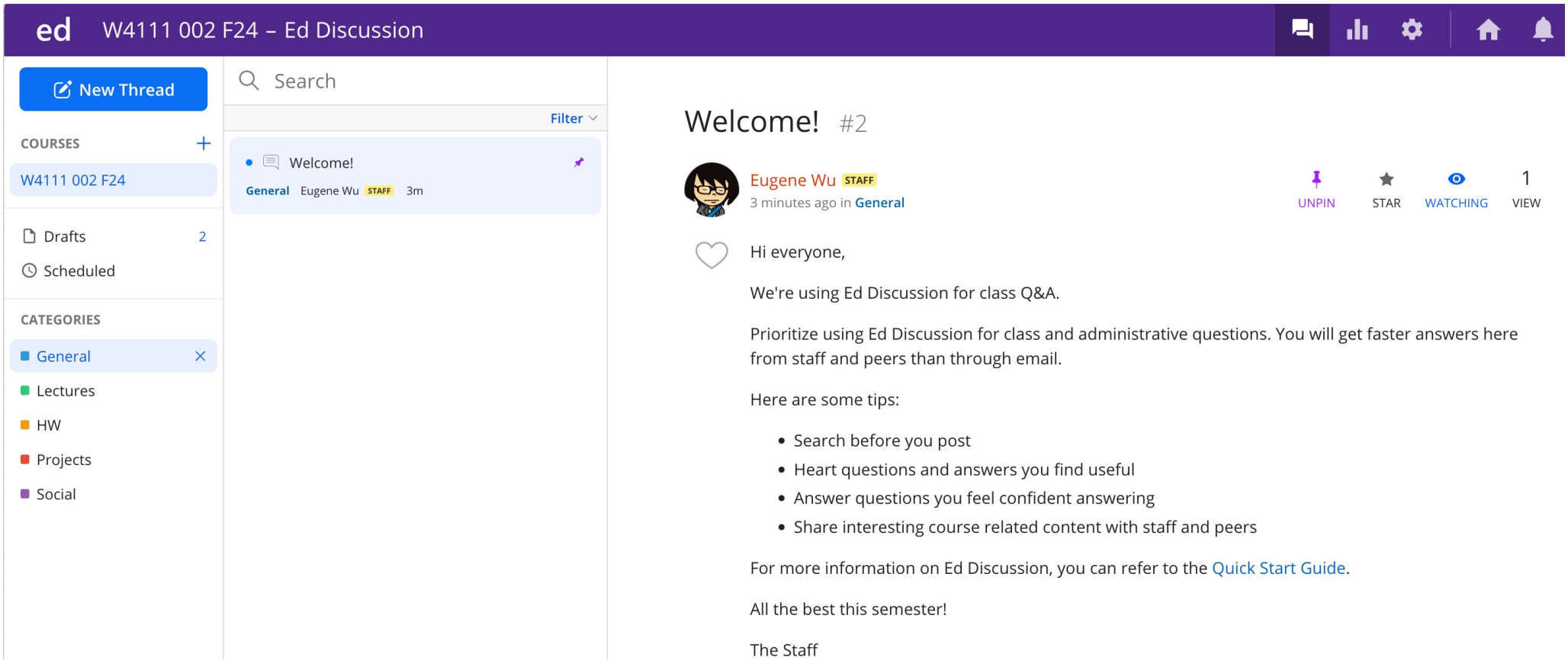
TBA

Your TAs

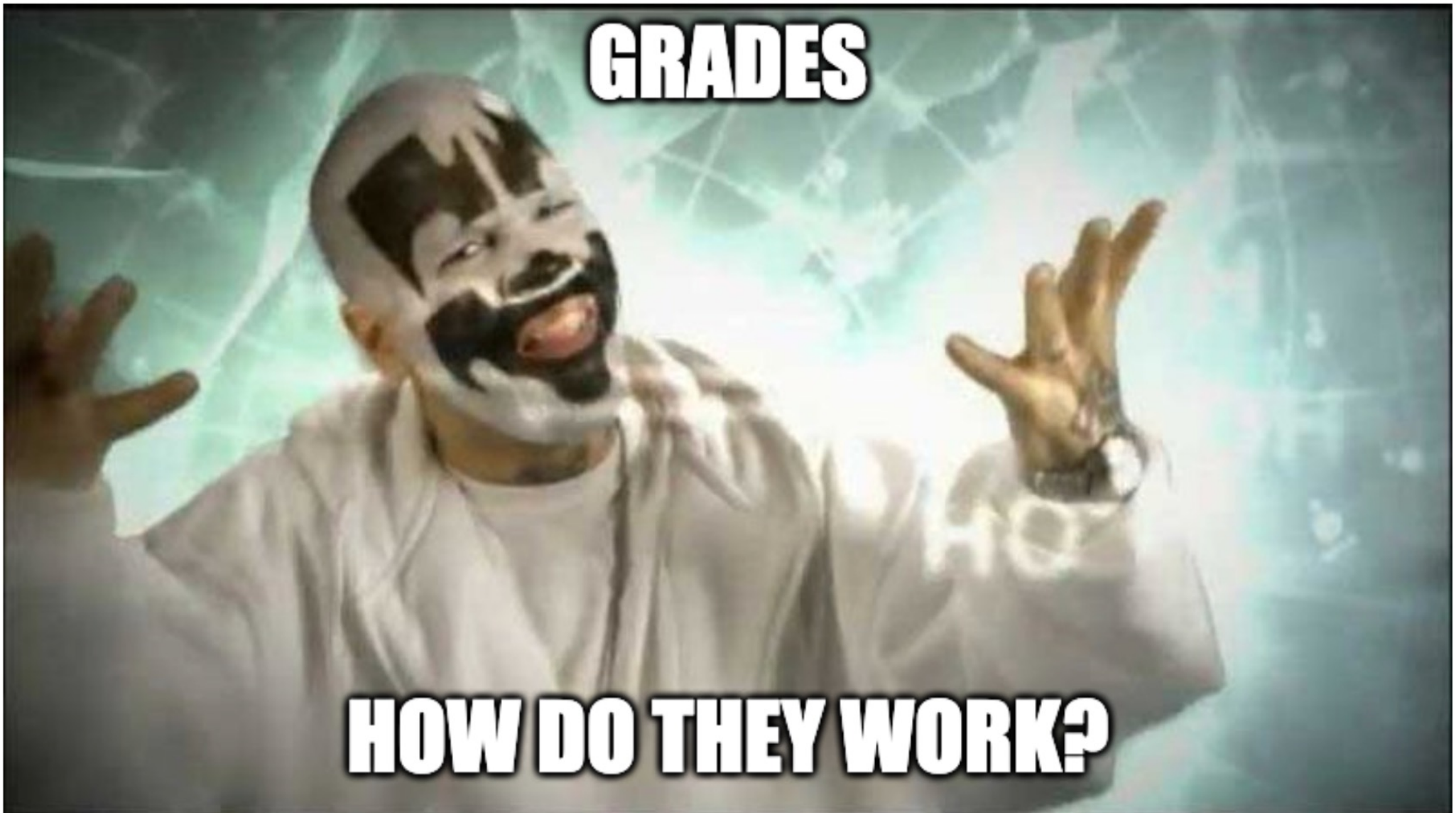
Office hours start next week.

- will be updated on course website
- Zoom links will be shared on discussion board

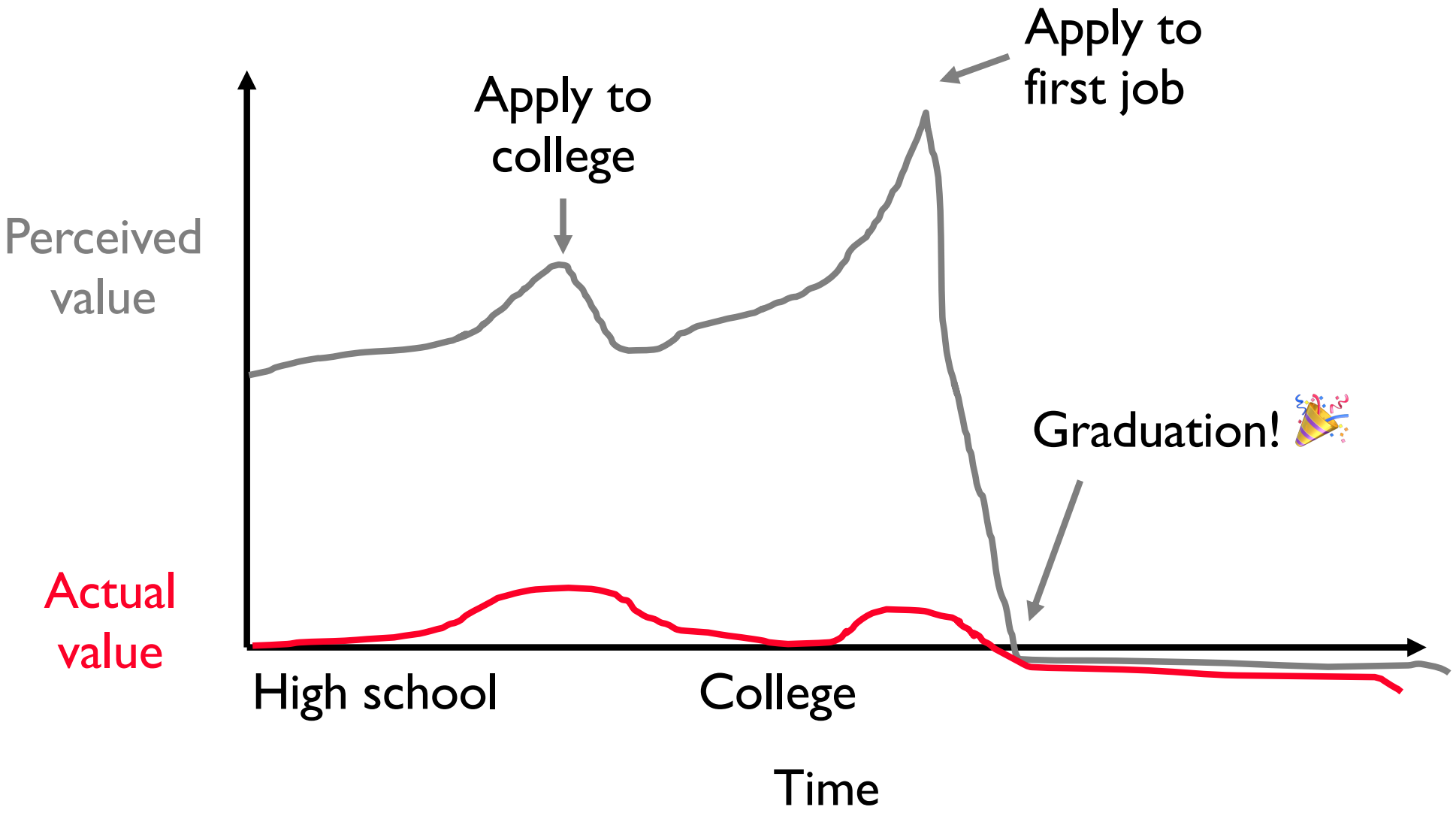
Discussion Board



Grades. How do they work?



Grades. How do they work?



Grading Information

Midterm 1	25%
Midterm 2	40%
HW	15% (4 HWs equally weighed)
Project 1	15%
Project 2	5%
Extra credit	variable

Median grade: historically B or slightly higher.

Exam Dates

Midterm 1	10/10	in person
Final	12/13	in person, cumulative

Makeup exams are not scheduled

Homework

Assignment will specify submission instructions.

No extensions or exceptions.

5 grace days for hws throughout the semester.

Can be applied to any assignment *unless otherwise specified*

After using all grace days, 25% grade deduction per day.

Don't need to tell us, staff will assign grace days in your favor

Check full details on web site under syllabus.

Projects (more details soon)

Two projects.

- Teams of two
- Run on cloud infrastructure
- Python & SQL

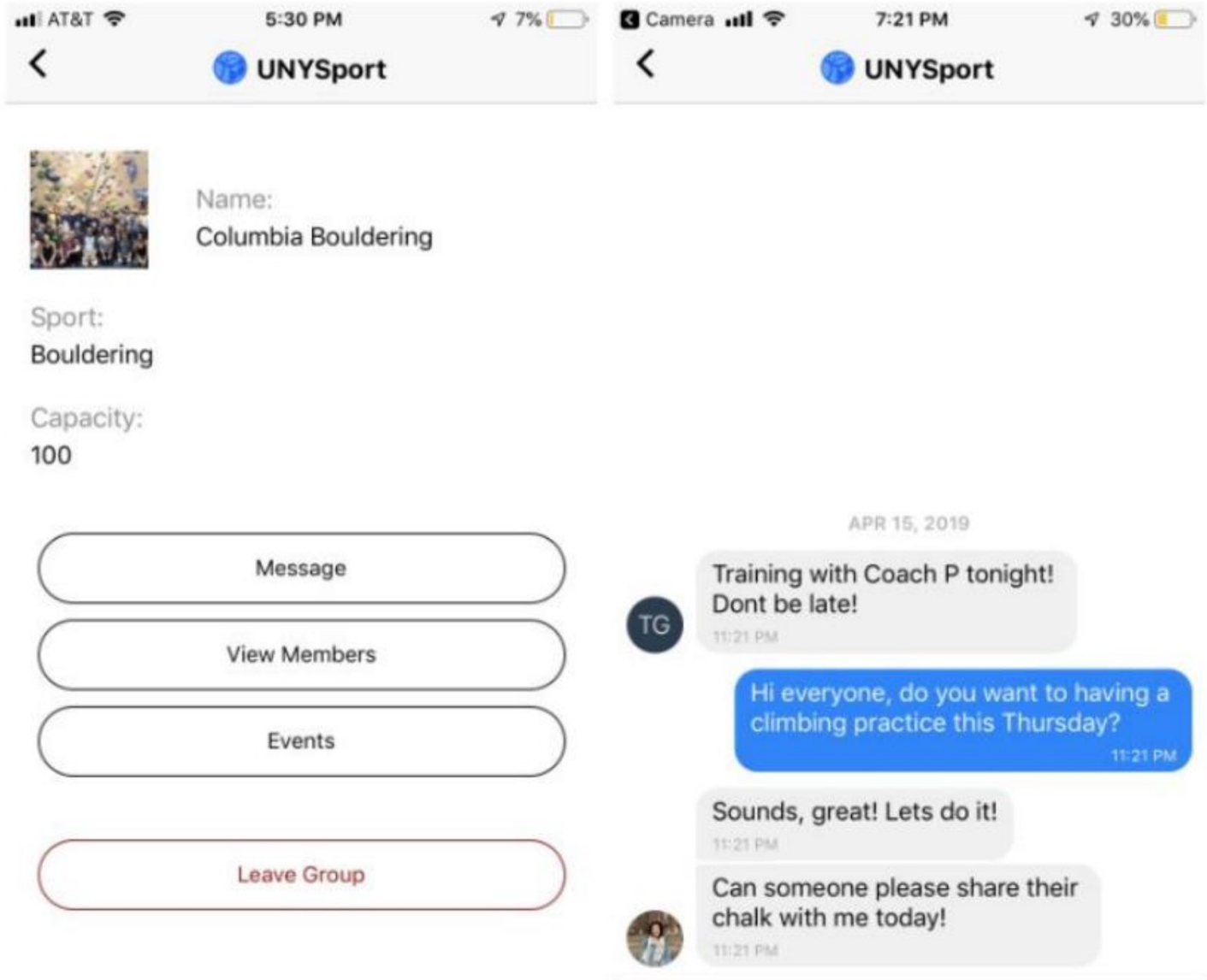
Project 1

- Model and build your own database web application
- Explore "traditional" relational database features.
- Non-programming option

Project 2

- Do cool things with DBMSes

Sports Community Mobile App



Course Match | Home | Department | Courses | Social | Search everything here

W4111 Introduction to databases

Department: Computer science

Description:

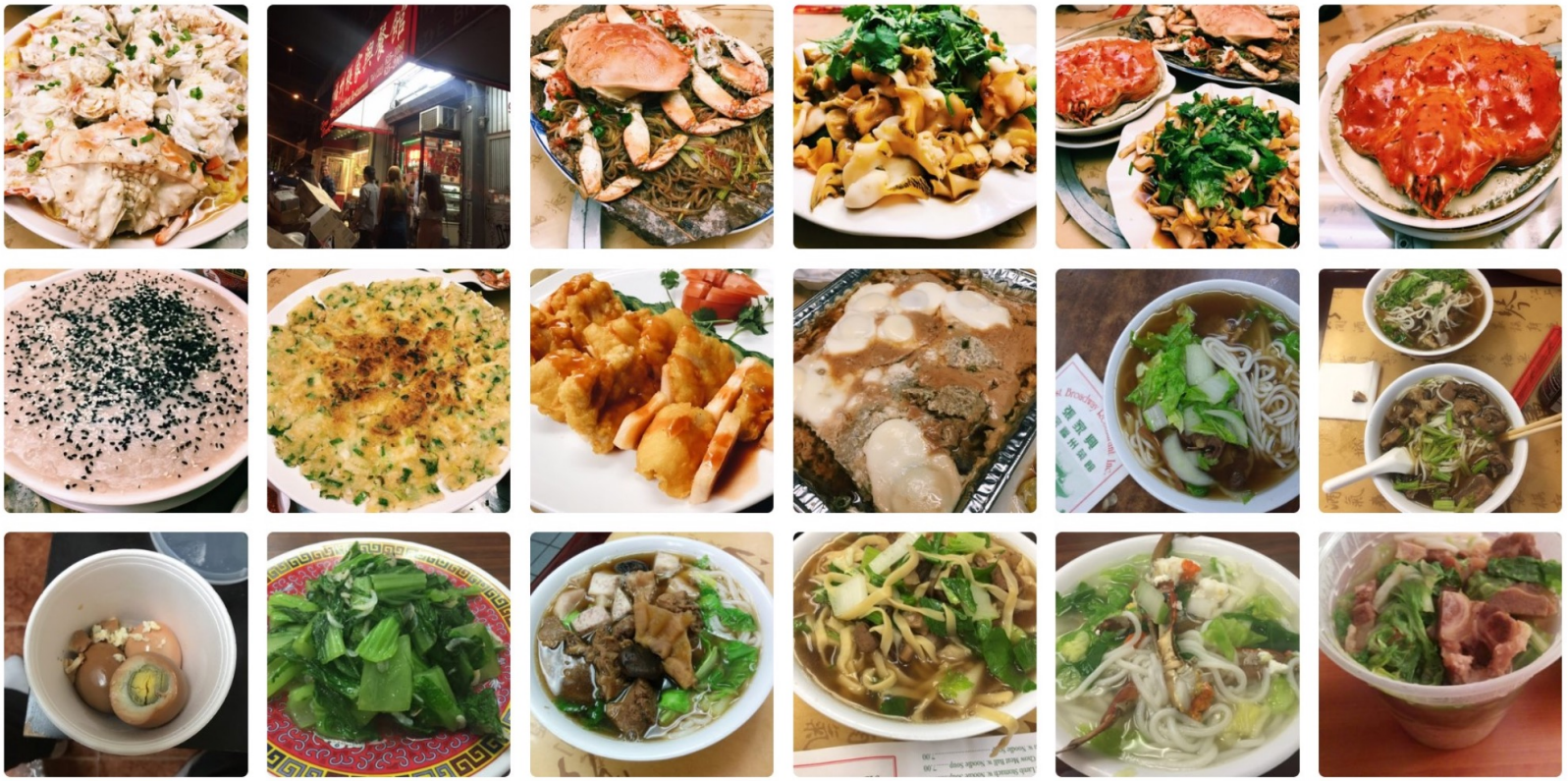
Prerequisites: (COMS W3134) or (COMS W3137) or (COMS W3136) and fluency in Java); or the instructor's permission. The fundamentals of database design and application development using databases: entity-relationship modeling, logical design of relational databases, relational data definition and manipulation languages, SQL, XML, query processing, physical database tuning, transaction processing, security. Programming projects are required.

Sections | Reviews

Instructor	Time	Day	Location	Year
Alexandros Biliris	13:10-15:40	Fri	To be announced	2019
Donald F. Ferguson	10:10-12:40	Fri	To be announced	2018
Eugene Wu	16:10-17:25	Tue, Thu	501 Northwest Corner	2018
Alexandros Biliris	16:10-18:40	Mon	750 Schapiro	2017
Eugene Wu	16:10-18:40	Mon	833 Seeley W. Mudd	2016
Alexandros Biliris	16:10-18:40	Mon	752 Schapiro	2016
Luis Gravano	16:10-18:40	Mon	753 Schapiro	2016

C-Food: Your guide to clean NYC Restaurants

East Broadway Restaurant



Borough: manhattan

Address: 94 East Broadway

Health Investigation Score: 41/50

Average User Rating: 3.20

Domino's

Projects (cont.)

3 grace days total for project parts 1 and 2.

No extensions or exceptions for project part 3 submission.

After using all grace days, 25% grade deduction per late day.

Check full details on web site.

Extra Credit

From Midterms, Projects, standalone, …

all added *after* the curve

Does NOT affect those that don't do extra credit

Collaboration Policy

Read Syllabus on course site for allowed conduct

CS Dept academic honesty policies
http://www.cs.columbia.edu/education/honesty

We will not tolerate *any* cheating

Collaboration Policy

Discussing lectures and course material strongly encouraged

Homework and exams are *individual*. No exceptions
Any libraries or code however minor must be disclosed.

Projects are done in *teams*; no collaboration between teams.

Contact the Professor Wu
right away if you have any questions or are falling behind.

Generative AI Tools

Coding assistants and large language models.

- Powerful tools that can "think and program for you"
- Tends to make up answers (hallucinate)

How will you know it's correct?

If it does your work, what is your salary for?

Learning and practicing fundamentals is necessary to verify *whether it is lying to you*

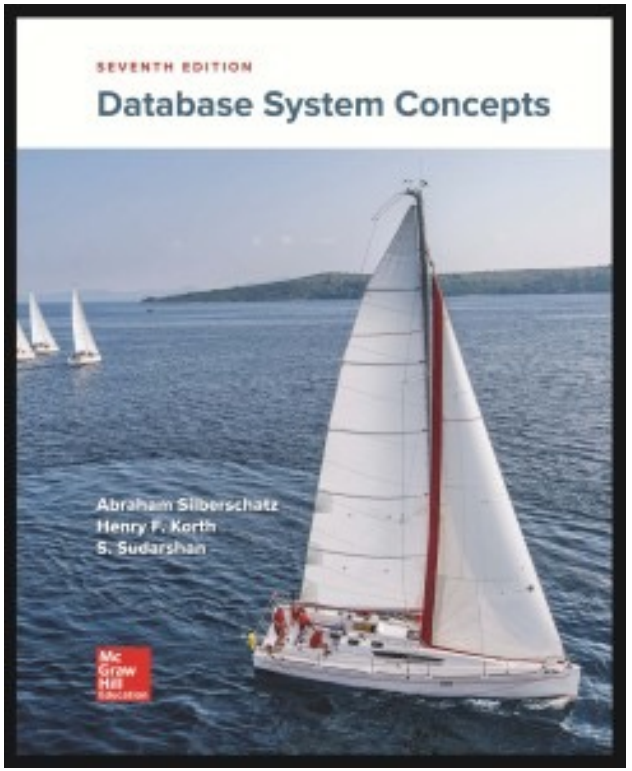
We will explicitly allow AI assistants for some assignments

Optional Textbook

Silberschatz et al.

Database System Concepts

7th ed



On-going Feedback

C O L U M B I A U N I V E R S I T Y C O M S W 4 1 1

INTRODUCTION TO DATABASES

Information

- Tues/Thurs
  8:40-9:55AM
  301 Uris Hall
  3 units
- Syllabus
- Ed Discussion
- Provide Feedback
- Course Github
- Course Gradescope

Overview

The goal of this class is two-fold. First, to introduce you to core database concepts (design, SQL) so that you too can build a billion dollar application. Second, to teach internals (e.g., physical database design, query optimization, transaction processing why queries may be running slowly/incorrectly. We will also discuss their relevance

The Data Management Seminar invites interesting database researchers and practi or on zoom (if available). We will announce these periodical

Announcements

Schedule

Data Courses at Columbia

COMS W4111 - Intro to Databases

Prerequisites: CS3137 or CS3134; fluency in Python

Intro to DBMSes

Data Models

Relational Algebra

SQL

Applications + SQL

Normalization

Peek at DBMS internals:

- Storage and indexing
- Query optimization
- Transaction Processing

COMS W4112-Database Sys. Impl.

Prerequisites: CS3137 or CS3134; fluency in Python

Components of a Database System in Detail

Storage Methods and Indexing

Query Processing and Optimization

Materialized Views

Transaction Processing and Recovery

Parallel & Distributed DBMSes

Performance Considerations Beyond Disk I/Os

COMS E6111-Advanced Databases

Prerequisites: CS4111; fluency in Java or Python

Information Retrieval
Information Extraction
Web Search
Data Mining
Data Warehousing, OLAP, Decision Support

COMS E6xxx-Graduate Seminars

Prerequisites: CS4111; fluency in Java or Python

6113 Database Research Topics

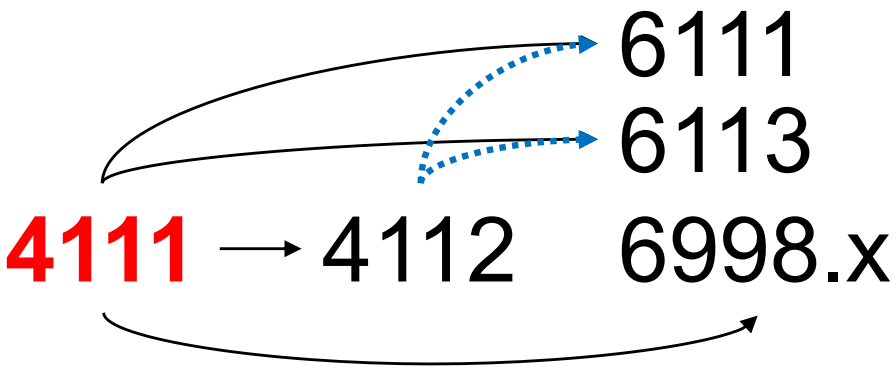
Topics vary e.g., modern databases, ML + Data

w6113.github.io

6998.002 Systems for Human Data Interaction

Topics combine HCI, visualization, and databases

columbiaviz.github.io


6111
6113
4111
4112
6998.x

Data Management at Columbia


Luis Gravano


Kenneth Ross


Eugene Wu


Kostis Kaffes


Mihalis Yannakakis

http://cudbg.github.io/

Borrowed material from
Prof. Gravano
Prof. Hellerstein (Cal)
Prof. Madden & Stonebraker (MIT)

w4111.github.io

DO HOMEWORK 0!